Genomics of sex determination in the mosquito *Aedes aegypti*

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of

Doctor of Philosophy

by

Joe Turner

November 2018

UNIVERSITY OF LIVERPOOL

Abstract

The mosquito *Aedes aegypti* is the principal vector of dengue, chikungunya and Zika arboviruses, all of which are growing public health concerns globally. Emerging technologies aim to control mosquito populations and limit disease transmission by releasing genetically modified mosquitoes, which could be improved by reliably manipulating sex determination. This relies on a strong understanding of the genetic basis of sex determination; however, little is known about this process in *Ae. aegypti* except that the male determining factor is located within a non-recombining, Y chromosome-like region on one chromosome called the M locus. In this thesis, I present experiments aimed at unravelling the nature of the *Ae. aegypti* M locus and analysing their impact on the evolution of mosquito sex chromosomes, with the intention of using the M locus as a target for sex-specific genetic modification.

Initially, I attempt integration of a fluorescence gene in male mosquitoes by targeting male-biased genomic sequences thought to be within or linked to the M locus with the RNA-guided endonuclease system CRISPR/Cas9. This involved the establishment of a laboratory mosquito line expressing the Cas9 endonuclease in the germline to facilitate effective transgenesis.

During this time, an M locus gene that acts as the sex determination switch, *Nix*, was discovered by other researchers. I designed experiments to further understand the nature of this gene by sequencing the region of the M locus in which it is situated. This resolved the gene structure of *Nix*, concluding that its intron is approximately 99 kb, which makes it one of the largest genes in the *Ae. aegypti* genome. The intron is enriched for repetitive DNA, suggesting that the inability of the M locus to recombine has led to the accumulation of transposable elements, in accordance with canonical models of sex chromosome evolution.

These findings were later expanded on, and I describe my role in analysing an improved *de novo* genome assembly for *Ae. aegypti*, determining nearly the full sequence of a 1.5 Mb M locus region, and mapping it physically to the chromosome. In addition, I undertook further research to examine the differential sex-biased coverage, structural variation, and repeat content of the M locus and the rest of the genome. I found that male-biased sequence and transposable elements have accumulated on the wider M chromosome, which may be indicative of it transitioning to a fully male-limited Y chromosome. This analysis illustrated the importance of high quality genomics data for studying mosquito sex determination, and could help to improve sex-specific targeting of genetic vector control strategies in the future.

Acknowledgements

First and foremost, I would like to thank Alistair Darby for his excellent guidance and encouragement throughout this PhD project. His patient mentorship helped develop my confidence as a researcher, and allowed me to pursue many opportunities I would not have otherwise attempted.

Though joining as a secondary supervisor midway through the PhD, Andrea Betancourt's enthusiastic support and attention to detail has been immensely helpful, and I am very thankful for her involvement.

I would like to thank previous project members Arjen van 't Hof and Ritesh Krishna, whose early contributions were enormously helpful, and who were always willing to follow up with information and advice after leaving the department.

Thank you to the other members of the Darby Lab: Sam W, Mark, Lauren, Stefany, Alex, Gwen; and former members, Fran and Dave. Our office has been a consistent place of lively chat and mutual aid, with everyone happy and willing to share knowledge and assist with analyses, that I can only hope is replicated in future research groups I might be part of. Special thanks to Stefany, Lauren and Mark for the many instances of cat-sitting! Thanks, too, must go to the members of CGR, especially Richard and Sam H for bioinformatics help and enjoyable pub trips; and John and Margaret for helping to carry out sequencing experiments.

I would also like to extend my thanks to my colleagues at Oxitec, whose help and guidance made my work there possible: in particular, Kelly Matzen, Sarah Scaife, Tabi Jenkins, Tarig Dafa'alla, David Navarro Paya, Pam Gray and Zoe Barnes.

Many thanks also to Ryan, Ed, Emily and Dylan for their support and friendship. I must also express my gratitude to my friends from undergraduate studies who were in Oxford during my placement, for their encouragement and camaraderie.

A massive thank you to my parents, who have always supported and believed in me, and the rest of the Turners, especially Helen and Fiona. I have also gained a second family during my time in Liverpool, in the Elliots, who have been unwaveringly enthusiastic and supportive, and I give them my warmest thanks.

Finally, thank you to my wonderful fiancée, Jess. Since we met you have been unfalteringly kind and understanding, cheering me on and welcoming me home at the end of the day. I could not be more grateful, and I'm happy that whatever adventure comes next, we will embark on it together

Contents

List of figures

List of tables

List of abbreviations

AaegL#	*Aedes aegypti* Liverpool strain reference genome version #
Aag2	*Aedes aegypti* cell line
AGWG	*Aedes aegypti* Genome Working Group
AmCyan	*Anemonia majano* cyan fluorescent protein
AWT	Asian wild type strain
BAC	Bacterial artificial chromosome
BAM	Binary Alignment Map
BED	Browser Extensible Data
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
BSA	Bovine serum albumin
Cas9	CRISPR-associated protein 9
cDNA	Complementary DNA
CGR	Centre for Genomic Research
CQ	Chromosome quotient
CRISPR	Clustered regularly interspersed palindromic repeats
crRNA	CRISPR RNA
DENV	Dengue virus
dNTPs	Deoxyribose nucleoside triphosphates
DsRed	*Discosoma* species red fluorescent protein
dsRNA	Double-stranded RNA
dsx	*doublesex*
EDTA	Ethylenediaminetetraacetic acid
EMBL-EBI	European Bioinformatics Institute
FISH	Fluorescent *in situ* hybridisation
fru	*fruitless*
$G_{\#}$	Post-injection generation number
gDNA	Genomic DNA
GFF	Generic Feature Format
GM	Genetic modification/genetically modified
HDR	Homology-directed repair
HGT	Horizontal gene transfer
LB	Lysogeny broth
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
LVP	Liverpool strain
LWT	Latin wild type strain
mRNA	Messenger RNA
Mya	Million years ago
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining

PacBio	Pacific Biosciences
PAM	Protospacer-adjacent motif
PCR	Polymerase chain reaction
piRNA	Piwi-interacting RNA
RIDL	Release of Insects carrying a Dominant Lethal
RNA-Seq	RNA sequencing (whole transcriptome shotgun sequencing)
RNAi	Interfering RNA
RPKM	Reads per kilobase million
rpm	Revolutions per minute
RT-PCR	Reverse transcription polymerase chain reaction
SAM	Sequence Alignment Map
SDS	Sodium dodecyl sulphate
sgRNA	Synthetic guide RNA
SINE	Short interspersed nuclear element
SIT	Sterile Insect Technique
smRNA	Small RNA
SMRT	Single molecule real time sequencing
SRA	Sequence Read Archive
TALEN	Transcription activator-like effector nuclease
tra	*transformer*
tracrRNA	trans-activating CRISPR RNA
UTR	Untranslated region
VCF	Variant Call Format
WGS	Whole genome sequencing
WHO	World Health Organization
WT	Wild type
ZIKV	Zika virus
ZMW	Zero-mode waveguide

Statement of work

The work presented here was conducted by the author, with the exception of contributions stated explicitly in the text and detailed below.

Chapter 2

Colleagues at the Centre for Genomic Research (CGR) and Oxitec Ltd performed sample material preparation, DNA extraction, library preparation, sequencing, and adaptor trimming on mosquito samples to generate genomic data. Ritesh Krishna performed bioinformatic analysis to identify target sequences. The Oxitec Molecular Team (Sarah Scaife, Tarig Dafa'alla, Caroline Phillips and Andrea Miles) PCR screened the target sequences in male and female mosquitoes.

Sarah Scaife designed plasmids containing the target sequences and primers for their synthesis. The CRISPR donor plasmids were built by the Oxitec Molecular Team (Sarah Scaife, Tabi Jenkins, Tarig Dafa'alla and Caroline Phillips).

David Navarro Paya assisted with some of the injections in the second CRISPR integration experiment, screened some of the mutant progeny and performed PCR experiments on the integration site.

Chapter 3

Kelly Matzen provided mosquito samples from Oxitec Ltd. Elizabeth Sutton and Alistair Darby commissioned the BAC library construction and prepared sample DNA. Arjen van 't Hof PCR screened the BAC library, identified and isolated the initial four BACs, performed BAC scaffolding, and extracted DNA.

Colleagues at the CGR performed library preparation and sequencing on BAC DNA. Alistair Darby assembled the DNA sequence. Ritesh Krishna mapped DNA and RNA data to the sequence. Alistair Darby and Ritesh Krishna drew Figures 3.3 and 3.4.

Chapter 4

The *Aedes aegypti* male reference genome assembly, AaegL5, was produced by the *Aedes* Genome Working Group (AGWG), established and coordinated by Ben Matthews and Leslie Voshall, of which the author was a member.

Atashi Sharma and Maria Sharakhova performed FISH using my M locus BACs, which determined the cytological location of the M locus. Zhijian Tu, Igor Sharakhov, Yang Wu, Alistair Darby and Alex Hastie collaborated on the work describing the M locus, including determining the location of the M locus in the genome assembly by alignment to the AaegL4 assembly and comparison with an optical map assembly.

Ritesh Krishna developed the initial bioinformatic analyses and MATLAB scripts to compare male and female coverage. Andrew Brantley Hall wrote the Chromosome Quotient pipeline.

Mark Whitehead developed the procedure for extracting high molecular weight DNA from insects. Colleagues at the CGR and at Edinburgh Genomics performed barcoding, library preparation, sequencing, and demultiplexing on the DNA for 10x linked read sequencing.

Mosquito rearing and DNA extraction from *Aedes albopictus* was conducted by colleagues at Oxitec Ltd. Sequencing was conducted by colleagues at the Oxford Genomics Centre in the Wellcome Trust Centre for Human Genetics.

Chapter 1 General Introduction

1.1 *Aedes aegypti*: a globally significant vector of disease

Aedes aegypti is the main vector of dengue virus, an increasingly important cause of tropical disease. The mosquito has spread with the rise of global trade and the rapid urbanisation and industrialisation in the tropics, providing breeding sites in the form of plastic containers and car tyres, which has led to a substantial increase in the global burden of dengue over the past century (Kyle and Harris, 2008). At least 2.5 billion people live in areas where they are at risk of dengue transmission from mosquitoes (Laughlin *et al.*, 2012) (Figure 1.1), and there are an estimated 390 million infections per year (Bhatt *et al.*, 2013). The incidence of dengue has increased significantly – by some estimates up to 30-fold – over the past 50 years (Pang *et al.*, 2017), and this is expected to rise further as *Aedes* mosquitoes continue to spread rapidly across the globe (Kraemer *et al.*, 2015). Most infections are asymptomatic but a significant minority develop as dengue fever, including around 500,000 annual cases of the more severe dengue haemorrhagic fever/dengue shock syndrome (Laughlin *et al.*, 2012). Furthermore, there is currently no effective vaccine that elicits immunity against all four viral serotypes (Simmons *et al.*, 2012). Recently, outbreaks of chikungunya and Zika viruses, both vectored primarily by *Ae. aegypti,* further highlight its public health importance (Musso *et al.*, 2015; Fauci and Morens, 2016). This has led to the development of methods for controlling the mosquito vector.

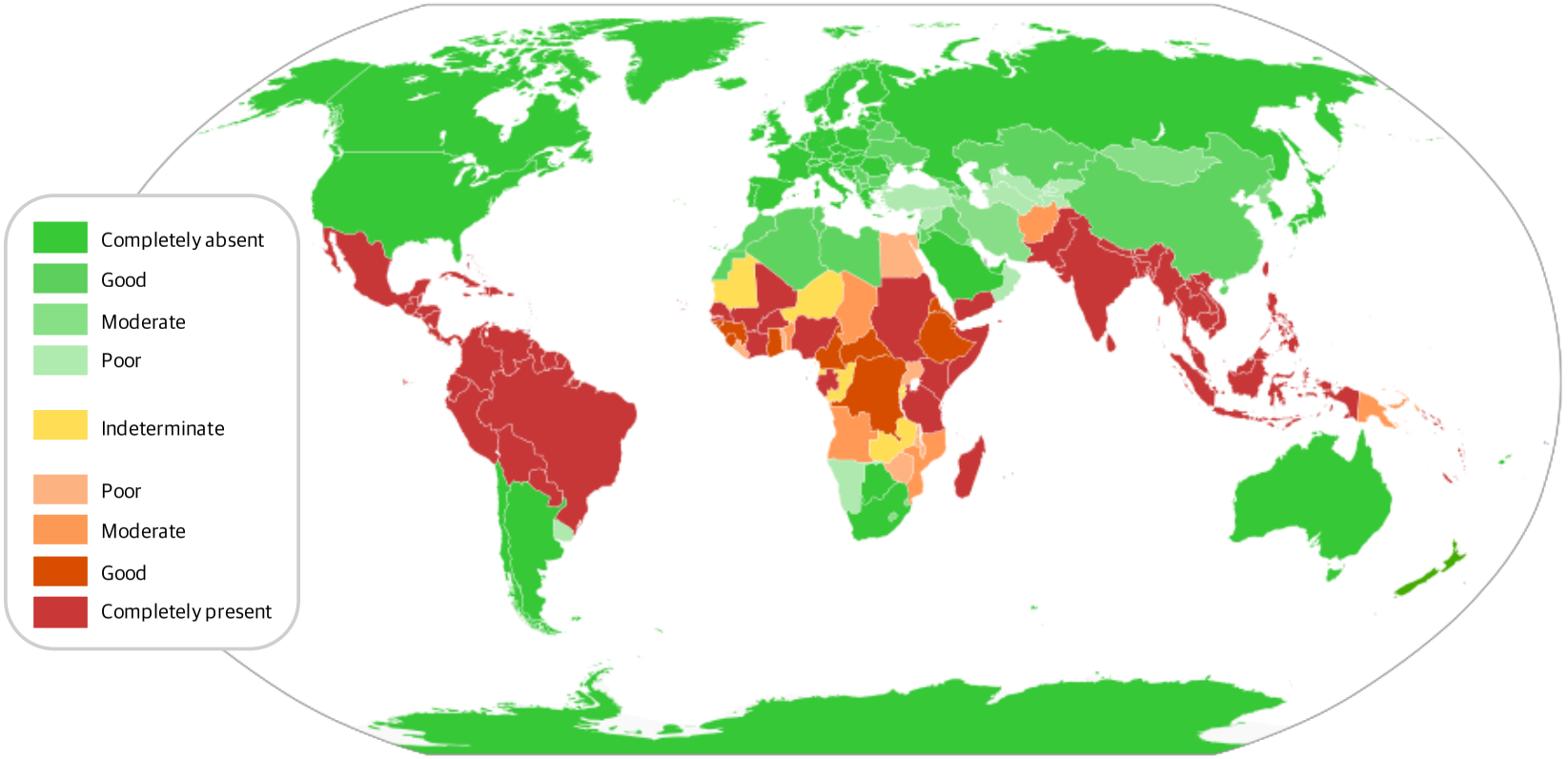


Figure 1.1 The global distribution of dengue. Countries are coloured by evidence consensus of dengue occurrence – a scoring system based on multiple weighted sources of evidence, including peer reviewed research, case data, data from health organisations like the World Health Organization (WHO), and other supplementary evidence such as news reports and prevalence of Ae. aegypti. *Intensity of red and green colours represent the strength of evidence that dengue was present or absent in each country. Figure drawn using data in Brady* et al. *(2012); map from Wikimedia Commons.*

1.2 Vector control strategies

1.2.1 Historical and contemporary control methods

One of the most commonly deployed strategies for reducing the population of mosquitoes is physical removal of breeding sites, sometimes called larval source management. *Ae. aegypti* larvae are notoriously well adapted to poor quality water, and their eggs can resist desiccation for months (Faull and Williams, 2015), which has allowed them to thrive in urban environments with prevalent sources of standing water like tyres, drums and jars (Hiscox *et al.*, 2013). Environmental management, often involving local governments, non-governmental organisations, and community engagement, is a common strategy to remove habitats in which mosquito larvae can breed. This can also take the form of large scale water storage and treatment practices (World Health Organization, 2009). Insecticides are also a widespread means to reduce mosquito populations. *Aedes* mosquitoes are day-feeding, preferring to bite humans in the evening, and therefore insecticide-treated bed nets, effective against malaria vectors, are of limited use. Instead, insecticides are applied directly to mosquito habitats. Larvicides are commonly applied to water sources such as discarded containers and tyres, and adulticides are applied by indoor residual spraying or space spraying (Kroeger *et al.*, 2006; World Health Organization, 2009; Paredes-Esquivel *et al.*, 2016).

The reliance on insecticides means that the evolution of resistance poses a challenge to future vector control programmes. *Ae. aegypti* populations resistant to all four of the main classes of insecticides (carbamates, organochlorines, organophosphates, and pyrethroids) have been observed in Africa, Asia and the Americas (Vontas *et al.*, 2012; Moyes *et al.*, 2017). Furthermore, the ecological impact of the heavy use of these insecticides must be considered. The combined environmental burden of chemical insecticides, mainly used to control agricultural pests but also disease vectors, is thought to be one factor in the recent alarming loss of insect biodiversity and resulting adverse effects on other organisms in the food web like insectivorous birds (Hallmann *et al.*, 2014; Hallmann *et al.*, 2017). Reducing the use of insecticides

in the context of the expanding global range of *Ae. aegypti* will therefore require the use of alternative vector control strategies, such as the release of genetically modified mosquitoes (Flores and O'Neill, 2018; Shaw and Catteruccia, 2018).

1.2.2 SIT and RIDL/"self-limiting" technology

The environmental impact of chemical insecticides and the emergence of resistance have created an interest in developing more sophisticated genetic techniques for vector control (Alphey *et al.*, 2013). One of these, the sterile insect technique (SIT), involves the release of insects that have been sterilised by radiation, which then mate with wild type insects without producing viable offspring, reducing the population of the insect (Benedict and Robinson, 2003; Schetelig and Wimmer, 2011; Alphey, 2014). However, radiation-induced sterility has the disadvantage of causing damage to somatic cells and symbiotic bacteria (Alphey, 2014). One promising variation on SIT is RIDL (Release of Insects carrying a Dominant Lethal), also known as "self-limiting" technology, which involves engineering a lethal trait inherited by the progeny of released genetically modified males. This lethality is repressible through provision of tetracycline in the larval water to allow rearing in captivity, but when absent in the wild, larvae die. These mosquitoes are therefore "genetically sterile" (Thomas *et al.*, 2000; Phuc *et al.*, 2007; Alphey *et al.*, 2013). Sustained releases of male RIDL mosquitoes have led to substantial reduction of the target population in previous contained laboratory (Wise de Valdez *et al.*, 2011) and field trials (Harris *et al.*, 2011; Lacroix *et al.*, 2012; Carvalho *et al.*, 2015).

1.2.3 Gene drive and other approaches

An alternative genetic method of control to SIT is gene drive, whereby a variety of synthetic selfish genetic elements could be utilised to spread genes in a mosquito population in violation of typical Mendelian inheritance (Burt, 2003). This approach could be used similarly to RIDL/"self-limiting" technology, but the desired genes would continue to spread even if they did not confer additional fitness (Sinkins and Gould, 2006; Macias *et al.*, 2017). Initially proposed decades ago, recent advances in

gene editing have made the prospects of this technology more realistic, although further improvements would be required for widespread deployment in the environment (Alphey, 2016). Various gene drive systems have been developed in *Anopheles*, showing promising success in spreading lethal genes that would crash natural populations; genes that confer resistance to the malaria parasite; and mutations causing complete sterility (Gantz *et al.*, 2015; Hammond *et al.*, 2016; Kyrou *et al.*, 2018). However none have been thoroughly tested in *Ae. aegypti*. Although theoretically targeted to a single species, the power of gene drives to eradicate whole populations has generated controversy, and the technology would need to be subject to stringent regulations to prevent unpredicted environmental damage (Oye *et al.*, 2014).

Another proposed mosquito control method involves the use of the widespread endosymbiotic bacterium *Wolbachia.* This symbiont is estimated to infect over 50% of arthropod species (Weinert *et al.*, 2015), and although *Ae. aegypti* is not one of these, when introduced artificially *Wolbachia* was found to reduce the ability of *Ae. aegypti* to transmit dengue and Zika viruses (Moreira *et al.*, 2009; Aliota *et al.*, 2016). Given that it can enhance its own transmission through cytoplasmic incompatibility, it is hoped that introducing *Wolbachia* into natural mosquito populations can induce the rapid spread of disease refractoriness (Jiggins, 2017), and some field trials have reported positive results (Hoffmann *et al.*, 2011). However, only the transmission of dengue serotype 2 is blocked, meaning that this strategy does not provide complete protection (Walker *et al.*, 2011).

1.2.4 Improving vector control through sex-specific targeting of genetic strategies

Genetic control techniques that involve the release of modified mosquitoes, including SIT, RIDL/"self-limiting", and gene drives, are constrained by the necessity of achieving reliable sex separation. Given that females bite humans and spread disease, it is crucial that females are eliminated from cohorts for mass release into the environment (Gilles *et al.*, 2014). This is sometimes done using size selection

based on the pupal sexual dimorphism in *Ae. aegypti*, however such a system is not completely stringent because pupae of the same sex can vary in size within a cohort (Papathanos *et al.*, 2009). Furthermore, mass rearing and separation of males and females is costly and time-consuming, which emphasises the potential benefits of implementing genetic sexing techniques. One technique utilises a ß2-tubulin promoter specific to the male gonads to express a fluorescence marker only in males, which can then be separated out (Catteruccia *et al.*, 2005; Smith *et al.*, 2007). In *An. gambiae*, gene-editing techniques employing homing endonucleases and the clustered regularly interspersed palindromic repeats (CRISPR) system were targeted towards the ribosomal DNA sequence on the X chromosome, resulting in a strongly male-biased sex ratio in the progeny (Galizi *et al.*, 2014; Galizi *et al.*, 2016). In *Ae.aegypti*, the self-limiting strain OX5034 contains a "flightless" construct that is mediated through sex-specific splicing of *doublesex*, resulting in female-only lethality (Fu *et al.*, 2007).

Besides the importance of strict sex separation for field release, sex-specific targeting is also useful because it could be exploited to increase the effectiveness of existing technologies. For instance, there is interest in developing transgenic mosquitoes with male-specific lethal effects (Hoang *et al.*, 2016; Sutton *et al.*, 2016), which could be limited to males more exclusively by inserting the constructs in a male-only genome region, mitigating the possibility of "leaky" expression in females. Whatever the technology used, robust and effective sex-targeting requires an understanding of sex determination systems in the relevant organisms.

1.3 Mosquito sex determination

1.3.1 Evolution of sex chromosomes

The existence of sex has long been considered one of the most fundamental mysteries in biology. Sexual reproduction – usually defined either very broadly as some form of genetic exchange between individuals, or somewhat more narrowly as some form of fusion of haploid cells produced through meiosis into a diploid zygote

(syngamy) – exists across all major groups of organisms. Sex in the former broad sense is ubiquitous even in prokaryotes, while meiotic sex is more typical of eukaryotes (Beukeboom and Perrin, 2014). While the fact that an organism's entire DNA is replicated in asexual reproduction while only 50% is in sexual reproduction may seem to make the former more advantageous, it is thought that the benefits resulting from the novel combinations of alleles generated during recombination in meiotic sex outweigh the cost of reduced transmission in certain contexts (Otto, 2009). Sex determination – the processes that underpin organisms developing into separate sexes which then produce gametes that fuse in reproduction – is achieved in an enormous variety of ways in eukaryotes, and can be induced by environmental or genetic factors (Bachtrog *et al.*, 2014).

Genetic sex determination was first recognised with the discovery of sex chromosomes, originally identified in the firebug *Pyrrhocoris apterus* when it was noticed that one element (named "X" because it was unknown) was only present in half of its sperm cells (Henking, 1891). Later work confirmed that sex chromosomes such as this X chromosome commonly underlie genetic sex determination (Beukeboom and Perrin, 2014). The most familiar system is XY (male heterogamety), where the male determination genes are on a male-limited Y chromosome which is morphologically different from the X. In the inverse of this (female heterogamety) the chromosomes are called Z and W; in species with haploid sex determination systems the chromosomes are called U and V (Coelho *et al.*, 2018). In organisms in which the male or female chromosome has been lost completely, so that the homogametic sex has two copies and the heterogametic sex has only one, the system is called XO or ZO.

The classical model for the evolution of differentiated sex chromosomes postulates that they evolve from autosomes with a sex-determination gene on one copy, leading to degeneration of the sequence around it (Charlesworth and Charlesworth, 2000; Charlesworth *et al.*, 2005). When a sex determination gene arises, it is advantageous for genes with functions that are beneficial in that sex to be inherited with it, but detrimental for them to be inherited in the other sex. Barriers to recombination at

these loci, such as inversions, are favoured by selection and increase in frequency in populations, causing the homologous chromosomes to become more isolated. The breakdown in recombination reduces the selection pressure against the proliferation of transposable elements (TEs) and loss-of-function mutations in sex-linked genes, a process known as Muller's ratchet, so repetitive DNA accumulates and pseudogenes form (Muller, 1964; Ellegren, 2011). The build-up of non-functional sequence can initially cause the chromosome with the sex determination gene to increase in size, however large deletions may occur (Bachtrog, 2013). Eventually, the chromosome with the sex determination gene becomes small, gene-poor and repeat-rich (Figure 1.2).

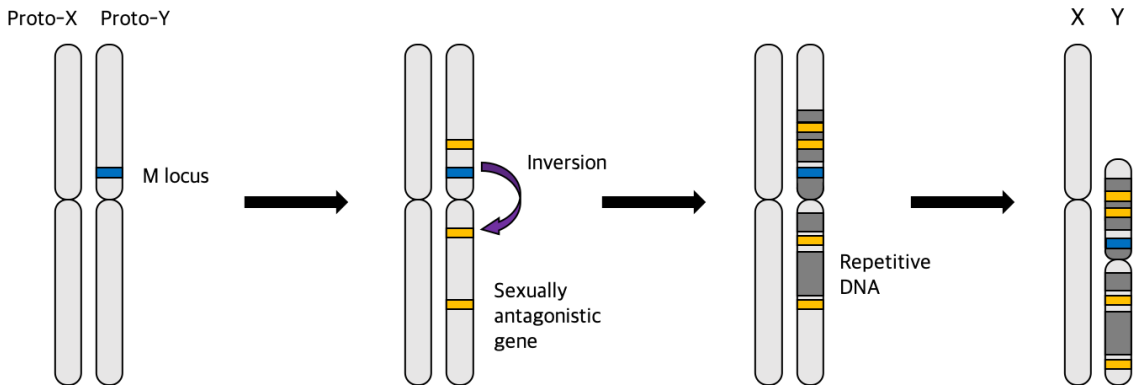


Figure 1.2 A model for the evolution of heteromorphic sex chromosomes. Initially, a sex determination gene (blue) arises on an autosome (e.g. an M locus on the proto-Y). Next, the accumulation of sexually antagonistic alleles (yellow) (e.g. those beneficial in males but detrimental in females) is favoured close to the sex determination locus. This leads to the reduction in recombination near these genes, for instance due to an inversion (purple), between the chromosome pairs. Repetitive DNA (dark grey) then accrues, for instance from introgression of transposable elements. Eventually, pseudongenisation and the expansion of repetitive DNA can lead to the loss of large portions of non-functional Y chromosome, resulting in differing physical sizes.

In insects, the most common systems are XY or XO male heterogamety, ZW or ZO female heterogamety, and haplodiploidy (Figure 1.3). The genes *doublesex* (*dsx*) and *fruitless* (*fru*), which are alternatively spliced according to upstream male or female signals on these sex chromosomes and regulate downstream sex development, are highly conserved across insects (Verhulst and van de Zande, 2015); however, the

signals upstream and downstream of these genes in the cascade are highly variable and evolve rapidly, taking different forms even in closely related species (Kaiser and Bachtrog, 2010; Vicoso and Bachtrog, 2013; Vicoso and Bachtrog, 2015).

The sex determination mechanism is most intensively studied in *Drosophila melanogaster,* which has an XY system but the Y chromosome does not control sex determination. Instead, the ratio of X chromosomes to autosomes controls the timing of the expression of the gene *Sex-lethal* (*Sxl*), such that males with one X translate a late-acting *Sxl* protein that causes the gene *transformer* (*tra*) to produce a truncated protein. This non-functional protein cannot initiate the female splice form of *dsx,* resulting in its male form and downstream male development (Charlesworth, 1996). In anopheline mosquitoes, XY chromosomes are the dominant form (Hall *et al.*, 2016; Bernardini *et al.*, 2017), with the upstream male-determining factor present on the Y chromosome (Criscione *et al.*, 2016; Krzywinska *et al.*, 2016). In the culicine mosquito lineage, many species have homomorphic chromosomes and the male-determining factor is present at an autosomal locus (Gilchrist and Haldane, 1947). These two mosquito clades diverged from a common ancestor over 200 Mya (Reidenbach *et al.*, 2009; Figure 1.4 shows the phylogeny and divergence times of the main Diptera species referred to in this thesis), making them an interesting case study for the evolution of sex chromosomes (Toups and Hahn, 2010).

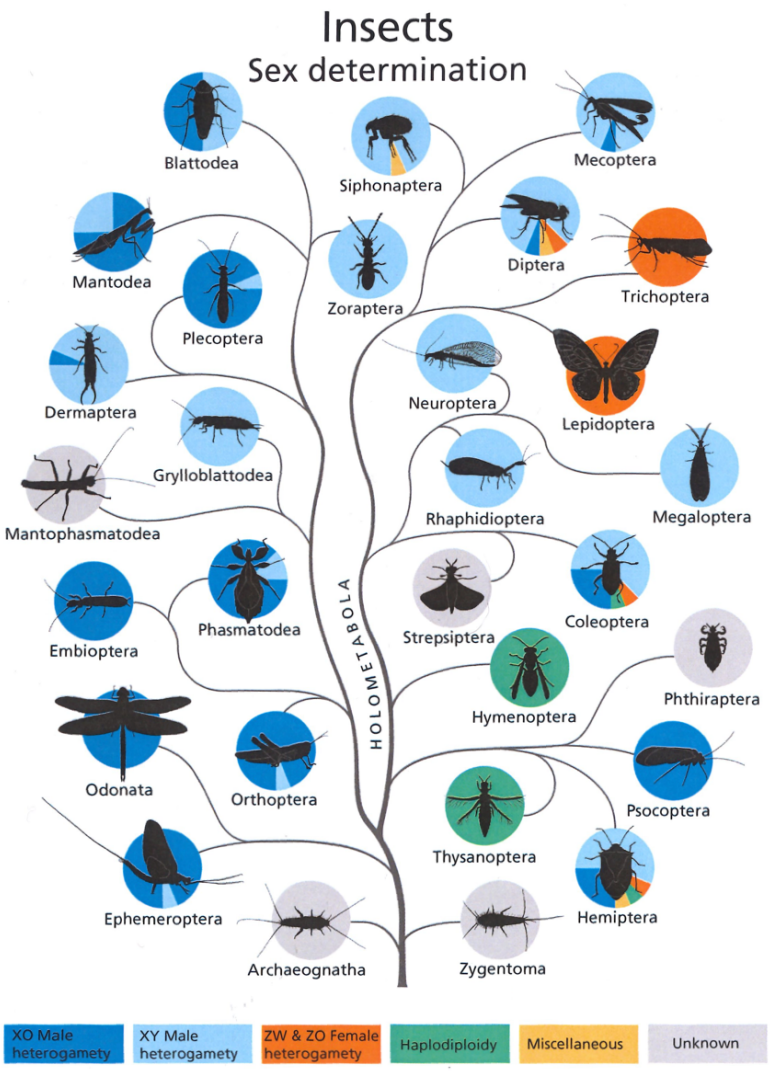


Figure 1.3 Phylogenetic tree of the different sex determination systems in the class Insecta. The coloured slices of the circles represent the approximate proportion of known species in each order with the corresponding system. The placement of the branches is based on the insect phylogeny in Misof et al. (2014); branch lengths do not indicate times since divergence. Figure from Beukeboom and Perrin (2014).

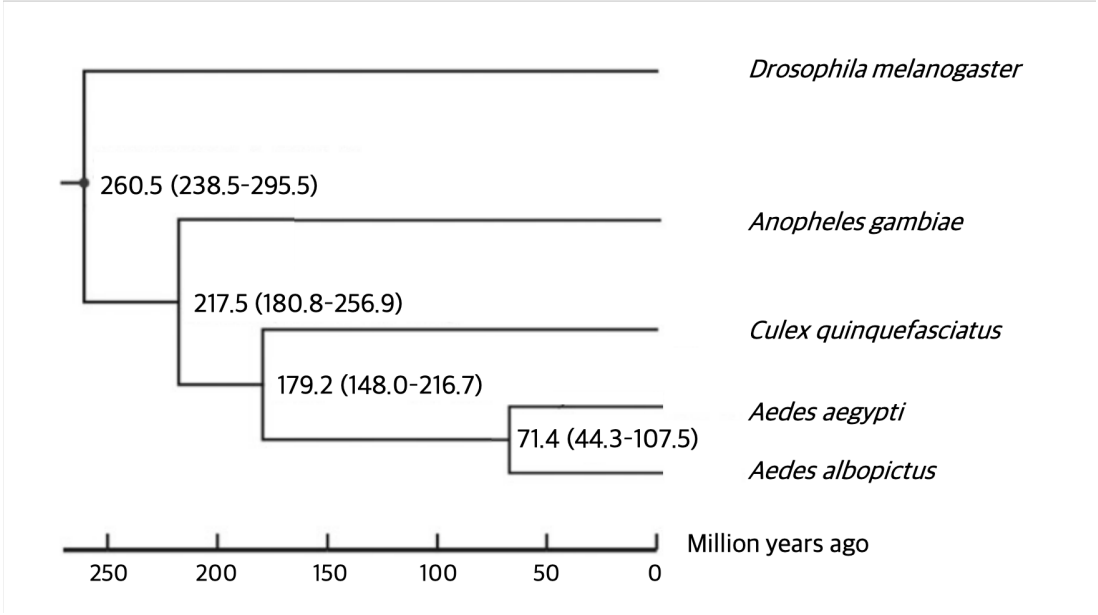


Figure 1.4 Phylogeny of four important mosquito species and the fruit fly D. melanogaster*, based on molecular clock analysis of 2,096 single-copy orthologues. Estimated divergence times are given for each branch, with standard error values in brackets. Figure redrawn using data from Chen* et al. *(2015).*

1.3.2 Sex determination in *Aedes aegypti*

Like other insects, the sex determination cascade in *Ae. aegypti* is mediated by the alternative splicing of *dsx* and *fru* (Salvemini *et al.*, 2011; Salvemini *et al.*, 2013). The upstream sex determination switch is not located on a Y chromosome; instead, male sex is initiated by an autosomal locus on one copy of chromosome 1, known as the M locus, outside of which the chromosome pairs are considered to be homomorphic. The corresponding locus on the other chromosome is referred to as the m, such that the chromosomes are sometimes called the M and m chromosomes. Recombination between these chromosome pairs is suppressed at this locus but recombination is thought to occur normally outside of it (Craig *et al.*, 1960; Hickey and Craig, 1966). It is not known if downstream transcription factors characterised in *Drosophila*, such as *transformer,* are shared by *Ae. aegypti* (Adelman and Tu, 2016; Biedler and Tu, 2016).

The nature of the M locus, such as the gene or genes responsible for initiating male development, has been a mystery due to the intractable nature of its genome. The *Ae. aegypti* genome is very large – approximately 1.3 Gb in comparison to the 278

Mb *An. gambiae* genome – and highly repetitive, with around 45% of the genome comprised of TEs (Nene *et al.*, 2007). The original assembly is fragmented, with a low proportion of the genome physically mapped due to the prevalence of these repetitive elements (Severson and Behura, 2012), although this was recently improved to 45% and 60% using FISH and genetic markers, respectively (Timoshevskiy *et al.*, 2014; Juneja *et al.*, 2014). Recent technological advances in genomics may be able to uncover the character of this enigmatic region.

1.4 Mosquito genomics

1.4.1 Insect genomics

The importance of genomics for informing research into insects has led to large scale efforts to produce high quality genomes. One initiative, Arthropod i5k, aimed to sequence the genomes of 5000 species of insects and other arthropods (Robinson *et al.*, 2011b). Even more ambitiously, the recently announced Earth BioGenome Project aspires to sequence all known eukaryotic species over the next 10 years (Lewin *et al.*, 2018).The availability of mosquito genomics data is valuable for its potential applications to disease control (Severson and Behura, 2012).
However, researching the genomics of the *Ae. aegypti* M locus is challenging. As mentioned above, the *Ae. aegypti* genome is fragmented with a low average contig size (contig N50 – defined as the length of the shortest contig, once all contigs have been ordered by length, that lies at 50% of the total genome size), having been sequenced using Sanger sequencing (Nene *et al.*, 2007), which often does not produce reads long enough to span TEs (Koren and Phillippy, 2015). In addition, pooled genomic DNA from male and female mosquitoes was used for the original genome project, meaning that only one quarter of the chromosome 1 sequences will be from a copy with the M locus, resulting in a consensus sequence that mostly derives from non-M sequences that likely obscures potential M candidates further (Hall *et al.*, 2014). Recently, more sophisticated technologies have been used to decipher the structure of sex chromosomes, such as in the fly *D. miranda* (Mahajan *et al.*, 2018),

and could be applied to *Ae. aegypti* to resolve some of the challenges of studying the M locus.

1.4.2 Single-molecule real-time sequencing

The use of long-read, single-molecule real-time (SMRT) sequencing is one way to overcome the problems associated with low-quality genomes (Metzker, 2010; Severson and Behura, 2012; Berlin *et al.*, 2015). The PacBio RS sequencer, released in 2011, is able to generate reads of much greater length than both the Sanger technology that originally sequenced the mosquito genome and next generation technology such as 454 and Illumina (Koren *et al.*, 2012). SMRT sequencing is based on visualisation of the incorporation of individual nucleotides into a DNA chain by DNA polymerase within a <100 nm well called a zero mode waveguide (ZMW), allowing the sequence of the growing DNA molecule to be deduced. (Eid *et al.*, 2009). Unlike other sequencing technologies, SMRT does not require amplification of the sample DNA, removing the need to deal with amplification-related artefacts (Niu *et al.*, 2010; Koren *et al.*, 2012), and is more reliable for sequencing highly GC-rich regions (Ross *et al.*, 2013). Crucially, the main advantage of PacBio is the length of reads, which at many kilobases can span repetitive elements and allow construction of highly contiguous genome assemblies (Koren *et al.*, 2012), and new assembly techniques have reduced the need to use high-accuracy short reads for error correction (Chin *et al.*, 2013). As a result, PacBio has led to a growing number of low-cost, high-accuracy de novo genome assemblies (Koren and Phillippy, 2015; Wee *et al.*, 2018).

1.4.3 Short read sequencing, 10x linked reads and other technologies

Although SMRT technologies such as PacBio have advantages over next generation sequencing (NGS) technologies based on short reads like 454 and Illumina for the purposes of genome assembly, especially with the declining need for their use for error correction in generating hybrid assemblies, these earlier tools still have important roles to play in analysing genomes and deciphering peculiar regions such

as the *Ae. aegypti* M locus. Newer machinery such as the Illumina HiSeq is capable of very high throughput and can allow low-cost whole genome resequencing at high coverage (Goodwin *et al.*, 2016). When a reference genome assembly is available, this can facilitate detection of polymorphisms in populations and the phasing of variants into discrete haplotypes (Davey *et al.*, 2011; Koren *et al.*, 2018), while sequencing and aligning male and female samples separately can enable the identification of sex-specific sequences (Hall *et al.*, 2013). Furthermore, these technologies are able to perform high-throughput RNA sequencing (RNA-Seq), which can be used for purposes such as transcriptome assembly, comparative expression, and population genomics analyses (De Wit *et al.*, 2012).
Recently, techniques have been developed to generate artificially long read data from short read sequence runs, such as 10x Genomics linked read sequencing. With this technology, high molecular weight DNA fragments undergo barcoding within droplets called GEMs, which are subsequently dissolved and the DNA sheared and sequenced with NGS platforms. Short reads with the same barcodes can then be linked together computationally, allowing parts of the same original long DNA strands to be detected (Goodwin *et al.*, 2016). Using only small amounts of input DNA, this method has been shown to be effective at identifying structural variants and phasing haplotypes (Zheng *et al.*, 2016), and can be useful for resolving large-scale genome architecture.

1.5 Thesis outline and aims

This thesis will attempt to characterise the sex determination genome region in *Ae. aegypti* known as the M locus, and apply this evidence to develop improved techniques for genetic vector control by, for instance, enhancing sex-specific targeting. Understanding the nature of the M locus, such as its content (genes and non-functional sequence) and how it regulates sex determination, as well as the wider genomic structure surrounding it and its population and evolutionary dynamics, may establish new avenues for genetic modification that will increase the

effectiveness of existing mosquito control strategies, and also deepen knowledge into the structure and function of mosquito sex chromosomes.

Chapter 2 describes experiments aimed at integrating synthetic DNA constructs at the M locus using the gene editing technology CRISPR/Cas9, and the engineering of a Cas9 endonuclease-expressing transgenic mosquito line. Chapter 3 details the sequencing and assembly of the primary M locus gene responsible for male development, recently discovered by other researchers, describing for the first time its full gene structure. Chapter 4 describes the author's role in an international collaboration to assemble an improved mosquito genome assembly containing the complete M locus sequence, which included deducing its physical location on the chromosome and analysing its male-specificity, as well as subsequent further research into the characteristics of the M locus and the wider male-limited chromosome.

Chapter 2 Targeted genome editing of the *Aedes aegypti* M locus using CRISPR/Cas9

2.1 Abstract

The bacterial immunity system CRISPR/Cas9 has rapidly developed into one of the most versatile and powerful tools for synthetic genome editing. By programming the Cas9 endonuclease to cut at a specific 20-nucleotide sequence, targeted double-stranded breaks can be generated, and with the provision of donor sequences it is possible to introduce relatively large DNA constructs into an organism's genome. This technology could be put to use in *Aedes aegypti*, an important arbovirus vector for which many genetic control strategies are being developed. In particular, the targeted modification of the sex-determining region known as the M locus could allow for sex-specific genetic engineering. In this chapter, CRISPR-mediated integration of a fluorescent marker gene was attempted into three sites hypothesised to be either within or linked to the M locus; male-specific fluorescence of the modified mosquitoes would then provide further evidence that these sequences are M-linked and demonstrate the potential for sex-specific modification. Due to low survival and lack of transgenesis, *piggyBac* transformation was used to generate a mosquito strain that expresses the Cas9 enzyme in the germline. Integration was reattempted at the known M locus gene *Nix*; yet despite higher survival of microinjected embryos, transgenesis was still unsuccessful. Overall, a single integration event occurred but was observable in both males and females, and therefore outside of the M locus. The results indicate that while CRISPR/Cas9 has been used to facilitate functional annotation of the *Ae. aegypti* genome, it is not especially suitable for investigating the content of the M locus. Although the experiments did not yield a successful method for male-specific CRISPR/Cas9 editing, future work may apply the tool to improve the genetic control of mosquito populations.

2.2 Introduction

2.2.1 Genetically modified insects

The prominence of insects as agricultural pests and vectors of disease has led to considerable efforts to develop targeted, non-insecticidal techniques for reducing their populations, such as artificial germline transformation using genetic modification technologies to produce mutant insects for mass release (O'Brochta and Handler, 2008; Criscione *et al.*, 2015). The mosquito *Aedes aegypti*, an important vector of arboviruses including dengue and Zika, is one of the more intensively studied insect species in this area (Alphey *et al.*, 2013). Early studies showed that microinjection of *Ae. aegypti* embryos with the *mariner* and *Hermes* transposable elements led to successful mutagenesis, demonstrating that stable transformation of the germline was possible (Coates *et al.*, 1998; Jasinskiene *et al.*, 1998). The *piggyBac* element, originally identified in the cabbage looper moth *Trichoplusia ni*, later emerged as an effective vector for insect transformation due to its relatively high efficiency in a variety of insects, inserting itself randomly at TTAA sequences (Handler, 2002). *piggyBac* has been used to design transgenic mosquito lines (Fu *et al.*, 2007; Labbé *et al.*, 2010), however this method does not allow precise editing of the genome. Elements that are able to recognise and target particular sequences such as TALENs and site-specific transgene integration systems have been used to transform *Ae. aegypti* (Nimmo *et al.*, 2006; Aryan *et al.*, 2013), yet these enzymes are still time-consuming to design and synthesise.

2.2.2 CRISPR/Cas9: A programmable genome engineering tool

Recently, a new technology promises to vastly improve the ability to manipulate genomes. The type II clustered regularly interspersed palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) system is becoming increasingly adopted as a gene-editing tool in a variety of organisms, including eukaryotes. Initially identified as an adaptive immunity-like system in prokaryotes, the

components used *in vitro* have been shown in recent years to induce site-specific cleavage of target DNA (Doudna and Charpentier, 2014). It functions by the Cas9 endonuclease being directed by two small RNAs – a CRISPR RNA (crRNA) and a trans-activating crRNA (tracrRNA) – to a particular sequence of DNA and catalysing the breakage of the DNA at this specific target (Jinek *et al.*, 2012). This breakage is repaired by the cellular DNA repair machinery via non-homologous end joining (NHEJ), which typically results in short nucleotide deletions or insertions (indels), causing frameshifts that can disrupt the target gene's function; however, incorporation of large DNA donor sequences can be achieved by exploiting the cells' homology-directed repair (HDR) pathway (Gratz *et al.*, 2014), potentially combined with the suppression of NHEJ (Basu *et al.*, 2015; Overcash *et al.*, 2015). Specificity to the target site is conferred by a 20-nucleotide spacer sequence in the crRNA, which pairs with the tracrRNA and directs Cas9 to the complementary DNA sequence (Figure 2.1). The target DNA must also have a three-nucleotide NGG protospacer-adjacent motif (PAM) flanking the 3' end of the complementary spacer sequence to facilitate target binding and cleavage (Hsu *et al.*, 2014). Thus, any 20-nucleotide sequence in an organism's genome adjacent to a PAM can be targeted with the right crRNA/tracrRNA, making the CRISPR/Cas9 system extremely easy to program for efficient gene editing compared to other existing techniques like TALENs (Jinek *et al.*, 2012). The two small RNAs can be combined into a single synthetic guide (sgRNA), reducing the number of components required for transformation (Doudna and Charpentier, 2014).

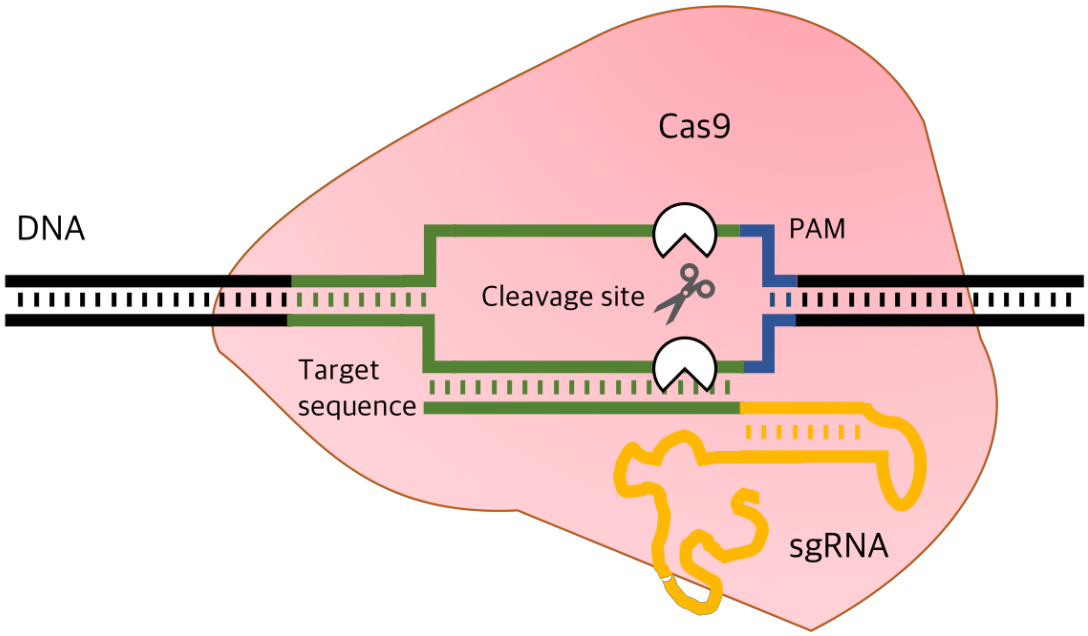


Figure 2.1 Schematic of DNA being cut by the CRISPR/Cas9 system. DNA is shown in black and the synthetic guide RNA (sgRNA) in yellow, with the target sequence of the DNA matching the sgRNA shown in green and the three-nucleotide NGG protospacer-adjacent motif (PAM) shown in blue. White markers show the cleavage sites in the target DNA sequence. Figure adapted from Ren et al. (2014).

Since its discovery and initial application, the sophistication and range of uses of the CRISPR/Cas9 system continues to expand (Barrangou and Doudna, 2016). As the structure of the Cas9 enzyme becomes elucidated in more detail, the amino acid sequence of particular domains can be tweaked to produce new variants with improved specificity or the ability to recognise alternative PAMs, leading to a reduction of off-target effects and a widening of targetable sequences (Kleinstiver *et al.*, 2015; Kleinstiver *et al.*, 2016; Chen *et al.*, 2017). CRISPR/Cas9 is also capable of binding RNA, suggesting that mRNA editing could be used for knockdown experiments (O'Connell *et al.*, 2014; Fonfara *et al.*, 2016; Lapinaite *et al.*, 2018). Although the majority of species that have been modified with CRISPR are model organisms, the list of organisms in which it has been attained continues to grow, and includes many insects.

2.2.3 CRISPR editing in insects

In insects, CRISPR has been used for multiple purposes, primarily either as a proof-of-concept to demonstrate effectiveness in a non-model species; as a method to investigate the function of particular genes through targeted knock-ins and knock-outs; and as a method to modify insects in predictable ways, for instance to engineer gene drives and other "synthetic genomic" functions (Bier *et al.*, 2018; Gantz and Akbari, 2018).

Initial application of the technique was successfully attempted multiple times in the model insect *Drosophila melanogaster* (Bassett *et al.*, 2013; Bassett and Liu, 2014; Gratz *et al.*, 2014; Port *et al.*, 2014; Ren *et al.*, 2014). These early studies established CRISPR/Cas9 as a robust method to modify insect genomes, requiring less costly and time-consuming optimisation than other gene editing procedures (Gantz and Akbari, 2018). They also showed that efficient mutagenesis can be achieved using a variety of concentrations of components and modes of delivery of the Cas9 enzyme (such as mRNA, recombinant protein, and germline expression) and the DNA donor (such as single-stranded oligonucleotides and double-stranded plasmids) (Sun *et al.*, 2017), and researchers continue to experiment with novel and synergistic approaches to further improve its effectiveness (Bier *et al.*, 2018).

Following *D. melanogaster*, CRISPR/Cas9 editing was accomplished and put to a range of uses in many other insect species, especially those considered economically important, including *Bombyx mori* (Wang *et al.*, 2013; Dong *et al.*, 2016), *Plutella xylostella* (Huang *et al.*, 2016), *Tribolium castaneum* (Gilles *et al.*, 2015), *Locusta migratoria* (Li *et al.*, 2016), *Apis mellifera* (Kohno *et al.*, 2016), and *Nasonia vitripennis* (Li *et al.*, 2017a). These uses included investigating the function of genes involved in sexual development in *Drosophila suzukii* (Li and Scott, 2016) and *Musca domestica* (Sharma *et al.*, 2017), indicating the potential of CRISPR/Cas9 for exploring – and possibly modifying – the genetic basis for sex determination in insects.

Amongst these insects, several mosquito species have been shown to be amenable to CRISPR/Cas9 editing. Efficient knock-outs of the eye pigmentation gene *white* have

been achieved in the malaria mosquitoes *Anopheles funestus, An. albimanus* and *An. coluzzii* (Li *et al.*, 2018), and in *An. gambiae* the immune gene *FREP1* was knocked out to reduce the susceptibility of the mosquito to the malaria parasite (Dong *et al.*, 2018). In the arbovirus vector *Culex quinquefasciatus*, CRISPR-mediated knock-out of the cytochrome P450 gene *CYP9M10* was used to validate its role in conferring pyrethroid resistance (Itokawa *et al.*, 2016). CRISPR/Cas9 mutagenesis was first reported in *Ae. aegypti* in 2015, and has since been developed and optimised further (Basu *et al.*, 2015; Dong *et al.*, 2015; Kistler *et al.*, 2015; Li *et al.*, 2017b), including using knock-ins and knock-outs to validate the function of the male-determining gene *Nix* (Hall *et al.*, 2015) and the dopamine receptor gene *Dop1*, which is involved in learning (Vinauger *et al.*, 2018). These successes suggest that CRISPR/Cas9 could be an important tool for controlling mosquito populations and the spread of disease.

2.2.4 Prospects for vector control using CRISPR/Cas9

The efficiency of CRISPR/Cas9 editing in a variety of insect species means it could be well suited to improving methods of biological control that utilise genetic modification, such as those in mosquitoes like *Ae. aegypti*. Existing strategies, for example those based on SIT and self-limiting transgenes, rely on the sustained release of male mosquitoes, and consequently the need to rear and separate out females is a significant drawback. Using CRISPR/Cas9 to generate mosquito lines with more males than females, or with sex-specific attributes, would be major improvements to such strategies (Bernardini *et al.*, 2014; Gilles *et al.*, 2014). For instance, in *D. melanogaster* it is possible to combine sgRNAs targeting a particular gene with ones targeting the female sterility allele ovo^{D1} so that the only progeny would have been successfully cut at this allele, thereby co-selecting and enriching for cuts at the desired target alleles (Ewen-Campen and Perrimon, 2018). A similar principle might be used in mosquitoes to bias transformants towards maleness.

Integrating transgenes directly into sex-specific parts of the chromosomes, such as the M locus, would allow expression to be narrowed to one sex and could be used to generate male-only rears for release into the field, or ensure male-only inheritance of the transgenes. Examples of sex-biased technologies have previously been developed, such as the genetic sexing and paternal effect self-limiting technologies, which utilise the alternative splicing of *doublesex* to trigger female-specific lethality and sperm-specific promoters to produce sterile progeny, respectively (Gong *et al.*, 2005; Fu *et al.*, 2007; Sutton *et al.*, 2016). However, more precise integration via CRISPR/Cas9 could increase the reliability of the sex-specific effects, for instance by preventing "leaky" expression of paternal effect constructs in females which reduces the fitness of transgenic lines. For instance, a paternal effect construct that uses the endonuclease FokI to inactivate sperm, causing infertility similar to that induced by SIT, can show strong off-target effects that can result in lower fitness in females (E. Sulston, data not shown). Inserting these paternal effect constructs in the M locus would ensure they are only inherited in males and prevent off-target effects in females. Specificity of CRISPR cutting also allows existing synthetic techniques to be combined to safeguard against the emergence of resistance to any single technique (Maselko *et al.*, 2018).

Another application of CRISPR/Cas9 to mosquito control is designing gene drives to distort typical Mendelian inheritance and spread desired mutations into wild populations. The framework for this idea has existed for many years and various approaches have been proposed, such as using underdominance and homing endonucleases (Burt, 2003; Windbichler *et al.*, 2011; Akbari *et al.*, 2013; Lambert *et al.*, 2018); however, the cost-effective and precise nature of CRISPR editing makes it a promising technique for engineering gene drives in mosquitoes (Adelman and Tu, 2016; Alphey, 2016; Macias *et al.*, 2017). Early attempts at constructing mosquito gene drive systems have experienced some success, managing to spread malaria resistance genes in *An. stephensi* and lethal genes to crash natural populations in *An. gambiae* (Gantz *et al.*, 2015; Hammond *et al.*, 2016). It may be possible to target sex-specific loci with CRISPR systems, similarly to the self-limiting

constructs described above. One promising direction is to directly use CRISPR/Cas9 to disrupt sex determining regions. In *An. gambiae*, an integrated Cas9 construct was used to cut the ribosomal DNA sequence on the X chromosome, which biased the sex ratio of progeny to up to 95% male (Galizi *et al.*, 2016), improving on a previously developed sex-ratio distortion system based on the homing endonuclease I-Ppol (Galizi *et al.*, 2014). Currently, this technology is likely to be more difficult in *Ae. aegypti* because less is known about what differentiates male and female chromosomes, meaning that further understanding of the M locus will be important for advancing such strategies.

2.2.5 Background and chapter aims

CRISPR/Cas9 gene editing has been used to explore the functional genomics of insects, and has also been proposed as a potential tool for the control of insects that act as crop pests or vectors of infectious diseases (Cui *et al.*, 2017; Sun *et al.*, 2017; Taning *et al.*, 2017; Gantz and Akbari, 2018). This chapter takes both approaches, exploring the use of CRISPR to both investigate the content of the *Ae. aegypti* M locus and to achieve sex-specific genetic engineering by attempting to integrate a fluorescent marker at this sex determining locus.

Prior to the work presented in this chapter, Ritesh Krishna developed a bioinformatic pipeline to calculate the differential breadth and depth of coverage of male and female DNA reads across the contigs in the reference VectorBase genome assembly AaegL3 (more information on this analysis and how it was subsequently developed by the author is given in Chapter 4.3.2). Of the ~36,000 genomic contigs, 35 were selected that showed a high degree of coverage of male reads but not female reads. The best candidates were identified by comparing the relative presence of each of these top 35 candidates in male and female DNA of five wild type backgrounds kept at Oxitec Ltd. Those that were found in males but not females across different wild-type backgrounds are most likely to be in linkage with the M locus. The three top candidates were contigs AAGE02035037.1(1-6260), AAGE02035965.1(1-4650) and AAGE02035016.1(1-6296).

The two primary aims of the work presented in this chapter are: 1) to determine whether the candidate sequences are truly male specific, which could allow further exploration of the *Ae. aegypti* M locus and the genetic basis for mosquito sex determination; and 2) to demonstrate CRISPR-mediated integration in *Ae. aegypti* mosquitoes as proof of principle so that this can be used in future to introduce constructs at specific locations in the genome, potentially advancing genetic strategies to control the spread of vector borne diseases.

2.3 Materials and methods

2.3.1 Mosquito rearing

2.3.1.1 Mosquito strains

Two principal laboratory strains of *Aedes aegypti* were used:

1. Asian wild type (AWT), also known as My1, originated in Jinjang, Kuala Lumpur, Malaysia and was colonised by the Institute of Medical Research, Kuala Lumpur, in the 1960s. The strain has been held at Oxitec since 2003, and was described in Lacroix *et al.* (2012).
2. Latin wild type (LWT) originated from 10 geographically separate locations in the state of Chiapas in southern Mexico, and was colonised in 2006 by combining equal numbers of individuals (approx. 50) from each location into a single colony to create a genetically diverse laboratory strain. The strain has been held at Oxitec since 2006 and was described in Wise de Valdez *et al.* (2010, 2011).

2.3.1.2 Egg hatching and larval rearing

All stages of the mosquito life cycle were reared under standard insectary conditions: 27°C ± 1°C temperature, 80% ± 10% relative humidity, and 12:12 h light to dark photoperiod.

Plastic deli pots of approximately 500ml capacity were used as hatch pots. Hatch pots were half-way filled with deionised water, egg papers were submerged, and the pots were covered in netting to avoid contamination between mosquito strains. To deoxygenate the water and stimulate synchronous hatching, the hatch pots were placed into a plastic vacuum desiccator and air was removed using a vacuum pump. The hatch pots were left for approximately 4 h, after which the pressure was equalised.

Hatched L1 larvae were transferred to 30 cm x 15 cm x 5 cm plastic rearing trays containing tap water, and reared at densities of 0.5 – 1 larva/ml. Larvae were fed on

TetraMin fish food (Tetra GmbH, Germany) as required. Trays were covered with netting to prevent cross-contamination between trays and eliminate the risk of potential escaped mosquitoes laying eggs in the water.

2.3.1.3 Sex separation of pupae

At rearing densities of 0.5 – 1 larva/ml, larvae normally begin to pupate 7 – 8 days after hatching, with males tending to pupate earlier than females. Pupae were removed from the trays by hand using a 3 ml pastette and transferred into plastic weigh boats.

Where necessary, pupae were sorted into males and females at this stage. In addition to size dimorphism (female pupae generally being slightly larger than males), *Ae. aegypti* pupae can reliably be sex-sorted by examining the dimorphism of their terminal segments (Figure 2.2). Water was mostly removed from the weigh boats, leaving a few millilitres, and pupae were examined under a dissection microscope, immobilised by placing the weigh boat on ice where required, and separated into males and females. In most cases for general rearing, such as colony maintenance and cages to produce eggs for microinjection, mosquitoes were combined at a 1:1 sex ratio. In other cases, the sex ratio is specified.

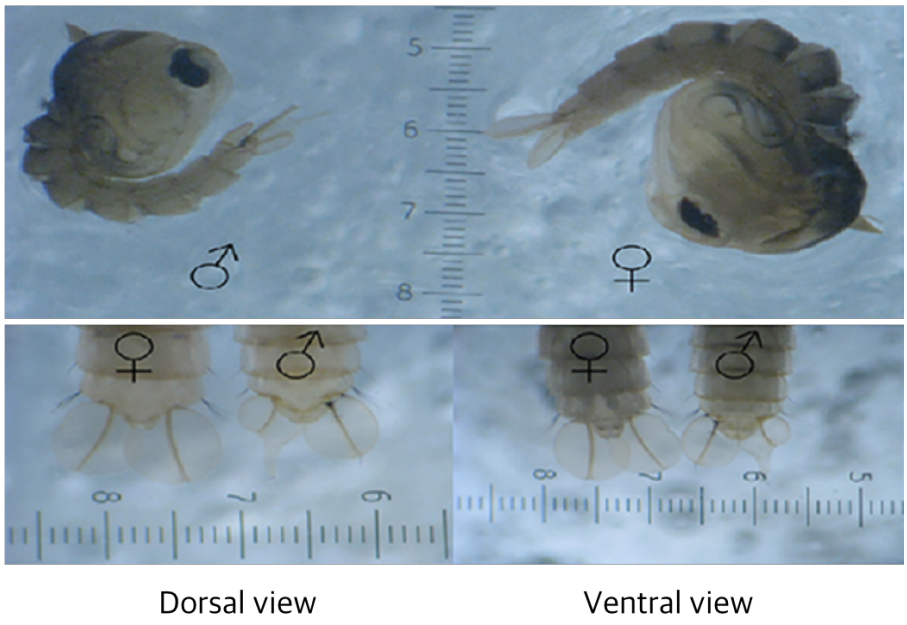


Figure 2.2 Sex dimorphism of Ae. aegypti *pupae. Males are generally smaller than females, and the morphology of the genitals on the terminal abdominal segment is noticeably differing morphology in males and females. Scale in mm. Figure adapted from Carvalho* et al. *(2014).*

2.3.1.4 Adult rearing and blood-feeding

Weigh boats containing pupae were transferred to 15 cm x 15 cm x 15 cm Bugdorm cages (Megaview, Taiwan). Pupae normally eclose into adults within 48 h. Adult mosquitoes were supplied with sucrose solution *ad libitum* in the form of 30 ml sugar-feeders containing a cotton wick, attached to the side of the cage with a plastic tag. The sucrose solution contained 0.2% Nipagin (Sigma-Aldrich, USA) to inhibit bacterial and fungal growth. Sugar feeders were changed after 7 days or the appearance of any microbial growth.

Adult mosquitoes were left for at least 3 days to mate and then blood-fed on defibrinated horse blood (TCS Bioscience, UK). Blood was placed on 5 cm x 5 cm or 10 cm x 10 cm metal plates and Parafilm was stretched over the plates so that it could be pierced by female mosquitoes' proboscises. These blood plates were placed on the top of the cages and heated to approximately 37°C by placing microwaved bean bags on top, which were reheated throughout the day to maintain the temperature. The cages were occasionally blown on gently to increase the levels of

carbon dioxide and encourage females to take a blood meal. Blood plates were usually left for 3 – 6 h to allow all mosquitoes to feed, and blood-feeding was repeated 1 – 2 times on subsequent days.

2.3.1.5 Egg collection

3 days after blood-feeding, strips of seed germination paper approximately 5 cm x 10 cm were labelled with the appropriate colony information (e.g. strain, generation, date), submerged in a small amount of water in weigh boats, and placed into the cages for females to lay eggs on. After 2 days, these egg papers were removed from the cages, drained of excess water, and stored in insectary conditions underneath netting to prevent contamination.

2.3.2 Preparation of injection components

2.3.2.1 Design of M locus CRISPR donor plasmids

For the first round of CRISPR knock-in experiments, three plasmids were designed at Oxitec Ltd. by Sarah Scaife that used between 800 bp-1kb of the candidate contig sequences (AAGE02035037.1, AAGE02035965.1 and AAGE02035016.1) as homology arms flanking the red fluorescence marker gene *DsRed2* followed by the sv40 polyadenylation signal to terminate transcription. The marker is under the control of a *ie1* promoter fused with homologous region 5 (*hr5*) enhancer for full-body expression (constructs OX5167-5169; Figure 2.3 shows a schematic diagram for OX5167).

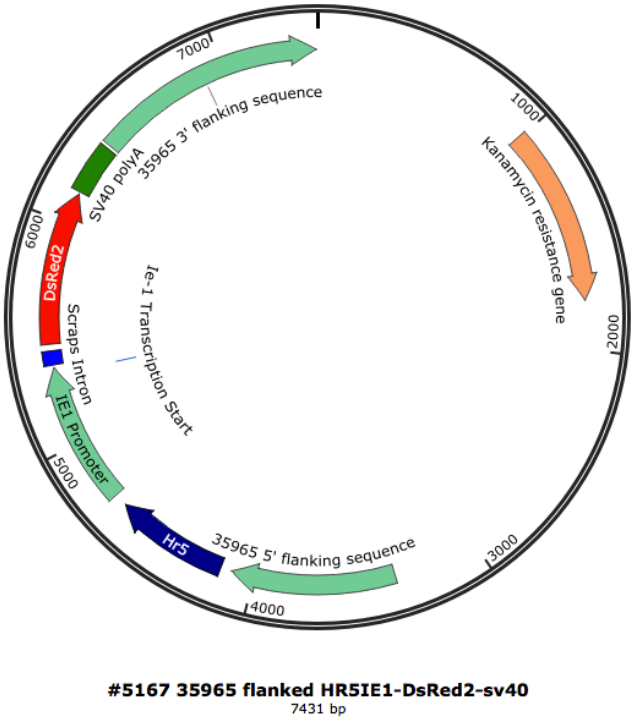


Figure 2.3 Schematic of one of the DNA plasmids used for CRISPR-mediated integration.

For the subsequent CRISPR knock-in experiments, Sarah Scaife designed a construct (OX5346; Figure 2.4) similar to the previous donor plasmids, but the flanking sequence surrounding *DsRed2* in this construct was the *Nix* gene and the surrounding sequence obtained from BAC library sequencing (described in Chapter 3). Additionally, the flanking sequences were extended because longer homologous sequences can increase the success of integration (S. Basu, personal communication).

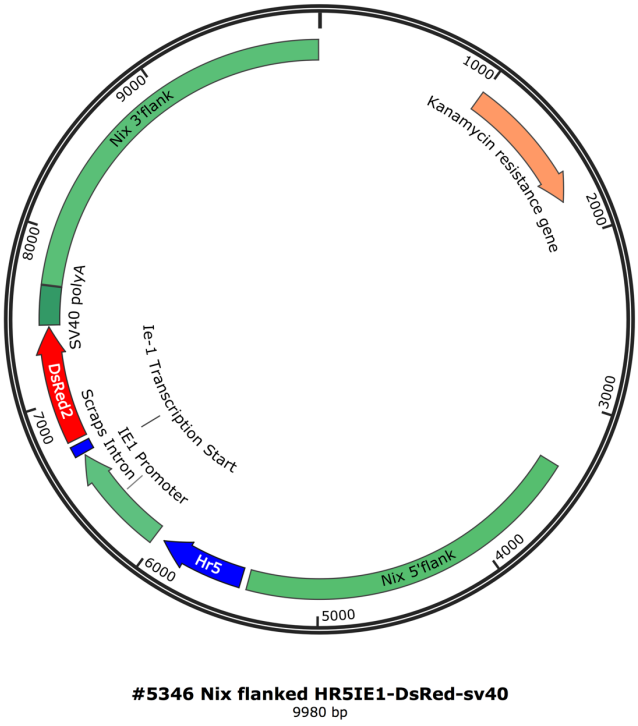


Figure 2.4 Schematic of the modified DNA plasmid used for subsequent CRISPR-mediated integration, incorporating the M locus gene Nix*.*

2.3.2.2 Design of piggyBac donor plasmid

A *piggyBac* construct (OX5226; Figure 2.5) was designed with the bacterial *cas9* gene under the control of the germline-specific *nanos* promoter (Adelman *et al.*, 2007), along with the *AmCyan* fluorescent marker gene.

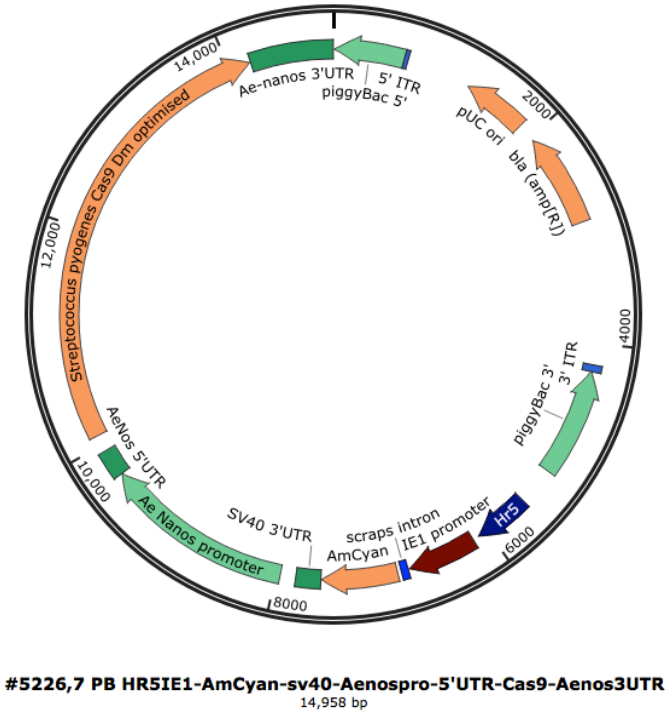


Figure 2.5 Schematic of the endogenous Cas9 piggyBac construct.

2.3.2.3 Construction of plasmids

Plasmids used for injection experiments were synthesised using standard transformation procedures. Individual components were amplified by PCR, using DNA containing the desired component sequences (such as mosquito genomic DNA or miniprep DNA of existing Oxitec Ltd. constructs) as templates. The components were joined and the plasmids cloned into *E. coli* and purified. The *piggyBac* construct OX5226 was built by the author and the CRISPR donor constructs OX5167-5169 and OX5346 were built by the Oxitec Molecular Team (Sarah Scaife, Tabi Jenkins, Tarig Dafa'alla and Caroline Phillips).

For the OX5226 *piggyBac* construct, four PCR reactions were run with Q5 High-Fidelity DNA polymerase (New England Biolabs, USA) in 50 µl volumes on a standard PCR thermocycling program to make the following components:

- Asc-Ae nanos promoter – 1725 bp
- Ae nanos 5' UTR – 252 bp
- Cas9 – 4256 bp

- Ae nanos 3' UTR-Xma – 763 bp

Additionally, a restriction digest was performed on the existing construct OX5053 with the enzymes AscI and XmaI (New England Biolabs) for 2 h at 37°C to produce an 8 kb vector backbone. The sizes of the amplicons were determined using gel electrophoresis and bands of the correct size were extracted and the DNA column purified using the GeneJet PCR Purification kit (ThermoFisher Scientific, USA) and quantified using a P300 Nanophotometer (Spectra, USA). 0.2 pmol of the combined components were incubated with 2x Gibson Assembly Master Mix (New England Biolabs) for 2 h at 50°C.

2 µl of a 1-in-4 dilution of the Gibson assembly product was added to XL10-Gold Ultracompetent Cells (Agilent, USA) mixed with β-mercapthanol and incubated on ice for 30 min. The cells were then heat shocked at 42°C for 30 s and returned to ice for 2 min. 200 µl transformation medium was added to the mix and incubated in a shaking incubator at 37°C for 1 h. The cells were then plated onto an ampicillin agar plate and incubated at 37°C overnight.

Colony PCR was conducted on the cultures to test for successful transformation (Bergkessel and Guthrie, 2013). Each colony on the ampicillin plate (~20) was picked and dipped into a well on a PCR plate, the reaction was run with BioTaq polymerase (PCR Biosystems, UK) to amplify a 511 bp product, and wells containing amplicons of the correct size were determined using gel electrophoresis. Eight wells that produced fragments of the correct size were selected and 10 µl of each well was added to 3 ml LB-Amp broth and incubated at 37°C, 250 rpm overnight. Between 1–1.5 ml of each culture was purified using the GeneJET Plasmid Miniprep Kit (Fermentas, USA) and a diagnostic digest was performed on 2 µl of each sample for 2 h at 37°C using FastDigest BglII and NheI restriction enzymes (ThermoFisher Scientific). The sizes of the fragments were determined using gel electrophoresis and the two samples with the correct sizes of 9.5 kb, 4 kb, 795 bp and 690 bp were Sanger sequenced by GATC Biotech (UK).

Sequences revealed that one sample was correct and one sample had a potential base substitution. 15 µl of the culture containing the correct sequence was added to 3 ml

LB-Amp and incubated at 37°C, 250 rpm for ~6 h and this preculture was then added to 250 ml LB-Amp and incubated overnight. The overnight culture was then purified using the EndoFree Plasmid Maxi Kit (Qiagen, Germany) and the same restriction digest assay and Sanger sequencing as used for the miniprep DNA was performed as described above. The plasmid sequence was determined to be correct and the maxiprep DNA was quantified and stored at -20°C.

2.3.2.4 Construction of guide RNAs

For the first round of CRISPR knock-in experiments, sgRNAs targeting the 3 candidate contigs were designed using standard CRISPR design tools optimised for *Ae. aegypti* (crispr.mit.edu), and any sgRNAs that would also target the donor plasmid were eliminated. In total, 18 sgRNAs – 6 targeting each candidate – were constructed. For the second CRISPR knock-in experiment, sgRNA sequences targeting the genes *kmo* and *Nix* were obtained from Basu *et al.* (2015) and Hall *et al.* (2015), respectively.

sgRNAs were synthesised using protocols adapted from Bassett *et al.* (2013). DNA templates were built using a no-template PCR, where the reverse primer is a common oligonucleotide sequence (SS1713) (Appendix 2.2) and the forward primers are the sgRNA sequences flanked upstream by a T7 promoter sequence and downstream by a sequence complementary to reverse primer, such that each primer had the structure

GAAATTAATACGACTCACTATA**[N]$_{20}$**GTTTTAGAGCTAGAAATAGC

where N_{20} is a 20-nucleotide sgRNA sequence identical to the genomic target. For sgRNA sequences that did not begin with 'GG', the two initial (5') bases were changed to 'GG' to achieve maximum efficiency of *in vitro* transcription from the T7 polymerase, such that the forward primer structure was

GAAATTAATACGACTCACTATA**GG[N]$_{18}$**GTTTTAGAGCTAGAAATAGC.

No-template PCR was performed with Q5 High-Fidelity DNA polymerase (New England Biolabs) in 100 µl volumes and run on a standard 35-cycle PCR program.

The PCR products were column purified using the GeneJet PCR Purification kit (ThermoFisher Scientific) and quantified using a nanophotometer.
500 ng of each purified DNA template was used for *in vitro* transcription using the Ambion MEGAshortscript T7 Transcription Kit (ThermoFisher Scientific): the reaction mixes were incubated at 37°C for 4 h and ammonium acetate solution was added to stop the reaction. The transcription products were purified by phenol-chloroform extraction: an equal volume of acid phenol-chloroform (125:21:1 phenol:chloroform:isoamyl alcohol for RNA extraction, ThermoFisher Scientific) was added to the reaction mixes, centrifuged for 5 min at 10,000 x *g*, and the aqueous layer was removed. An equal volume of chloroform (ThermoFisher Scientific) was added to the aqueous phases and centrifuged under the same conditions, and the aqueous layers removed. RNA was obtained from the aqueous phases using alcohol precipitation: 2 x volume of 100% ethanol was added and the mixtures incubated at -20°C for 15 min, then centrifuged at 4°C for 15 min at 10,000 x *g*. The supernatant from each mixture was removed and the pellets were resuspended in nuclease-free water. The sgRNA suspensions were quantified and 1 µg/µl aliquots were stored at -80°C.

2.3.2.5 Construction of dsRNA

Primer sequences to synthesise the DNA template for dsRNA targeting *ku70* were obtained from Basu *et al.* (2015) (SS2081 and SS2082) (Appendix 2.2). No-template PCR was carried out according to the same protocol for the sgRNAs above. dsRNA was prepared using the Ambion MEGAscript RNAi Kit (ThermoFisher Scientific) and purified using the Ambion MEGAclear Transcription Clean-Up Kit (ThermoFisher Scientific). The dsRNA suspension was quantified and 500 ng/µl aliquots were stored at -80°C.

2.3.2.6 In vitro test of CRISPR activity

The ability of the 18 sgRNAs prepared for the first round of CRISPR knock-in experiments to cut the correct targets was tested with an *in vitro* incubation assay.

DNA templates for each of the three candidate sequences, each containing the six respective sgRNA cutting sites, were synthesised by PCR using BioTaq polymerase (PCR Biosystems) and a template of LWT genomic DNA. The PCR products were examined using gel electrophoresis, and bands of the correct size were cut out of the agarose gel and purified using the QIAquick PCR Purification Kit (Qiagen).

10 µl reaction tubes were prepared containing 250 ng of each sgRNA with 100 ng of its respective target DNA, along with 350 ng recombinant Cas9 protein from *S. pyogenes* (P&A Biotech. China), NEBuffer 3 (New England Biolabs) and 0.1 µg/µl BSA. The mixtures were incubated for 1 h at 37°C, 1 µl of 4 µg/µl RNase A was added and incubated for a further 15 min, then a SDS stop solution (1.2% SDS, 30% glycerol, 250mM EDTA pH 8) was added and incubated for a further 15 min.

The products were examined using gel electrophoresis (an example given in Figure 2.6). For each of the 3 targets, the 2 most effective sgRNAs targeting the 3' and 5' strands (i.e. those for which the smaller fragment bands were brighter and the original DNA template band was fainter) were selected to inject.

The sgRNAs prepared for the second round of CRISPR knock-in experiments were not assayed because their reported success in published studies was deemed to be sufficient evidence of their activity.

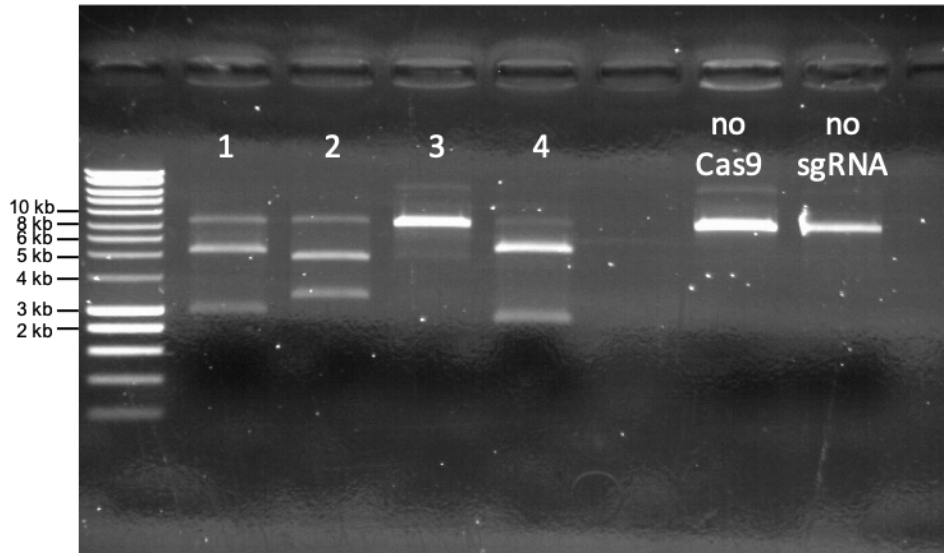


Figure 2.6 Results of an in vitro *test of CRISPR activity. A DNA fragment containing the target sites was incubated with four sgRNAs (1-4, corresponding to sgRNAs 2127-2130), along with controls containing no Cas9 enzyme or no sgRNAs. Assays 1, 2 and 4 show that the target has been cut successfully.*

2.3.2.7 Preparation of injection mix

An evaluation of the available literature on CRISPR/Cas9 modification of *Ae. aegypti* was conducted to determine the optimal relative proportions of components to inject. For instance, Dong *et al.* (2015) used 1 µg/µL Cas9 mRNA and 50 ng/µL of each sgRNA; Basu *et al.* (2015) and Hall *et al.* (2015) used 600 ng/µL Cas9 mRNA or recombinant protein and 100 ng/µL of each sgRNA; and Kistler *et al.* (2015) used 300 ng/µL of Cas9 protein and 40 ng/µL of each sgRNA, along with 500 ng/µL of dsDNA plasmid donor. The range here suggests that successful transformation with CRISPR/Cas9 can be achieved without the need for especially precise ratios. Kistler *et al.* (2015) state that recombinant Cas9 protein produces a higher rate and more reproducible mutagenesis (up to 5-10x higher) than mRNA, and results in better survival of embryos. This may be because there is no delay while the mRNA is translated, allowing the sgRNAs to form stable complexes with Cas9 prior to injection (Jinek *et al.*, 2014). They also found that increasing the sgRNA concentration did not significantly affect the rate of mutagenesis above 40 ng/µL.

A final composition was set at 50 ng/µL of each sgRNA, 350 ng/µL Cas9, 500 ng/µL donor plasmid DNA and 67 ng/µL *ku70* dsRNA. For the *piggyBac* transformations, the composition was 300 ng/µl plasmid DNA with 700 ng/µl helper mRNA. Table 2.1 shows the relative concentrations of the constituent components in the mixes for each construct. Components were combined with *Ae. aegypti* injection buffer (0.1 mM $NaPO_4$, 5 mM KCl, pH 6.8) and aliquots were stored at -80°C.

On injection days, aliquots of injection mix were thawed on ice and spun using a microfuge.

Table 2.1 Concentrations of the injection mix components for the germline transformation experiments with the three M locus candidate constructs, the piggyBac *construct, and the* Nix *construct.*

	OX5167-5169	**OX5226**	**OX5346**
Component			
Donor plasmid DNA	500 ng/µl	300 ng/µl	500 ng/µl
sgRNA (each)	50 ng/µl	–	50 ng/µl
ku70 dsRNA	67 ng/µl	–	67 ng/µl
Cas9 recombinant protein	350 ng/µl	–	0 or 350 ng/µl
piggyBac helper mRNA	–	700 ng/µl	–

2.3.3 Germline transformation

Germline modification of *Aedes aegypti* was conducted according to a procedure for microinjection of embryos, adapted from similar protocols (Jasinskiene *et al.*, 1998; Lobo *et al.*, 2006).

2.3.3.1 Egg collection and preparation

Mosquitoes were reared according to the methods described above, at the lower end of larval density to allow females to grow as large as possible and produce embryos that are larger and thus easier to manipulate and inject. Blood-feeding was performed 3 days before injection days.

On injection days, circles of Whatman grade 3 filter paper were moistened by placing them on damp cotton wool in Petri dishes and the dishes were placed inside the blood-fed cages. Cages were placed in the dark for 45 min – 1 h to stimulate synchronous oviposition, after which the egg papers were removed.

After oviposition, mosquito embryos gradually melanise and the chorion hardens so that they darken from white to black in 1 – 2 h in insectary conditions. Mid-to-dark grey embryos were chosen for injection because white embryos are more easily damaged by manipulation, while darker, mature embryos have a harder chorion that can cause the injection needle to break, and the success of transformation declines after nuclei begin to divide (Figure 2.7). Embryos were transferred to new moistened filter paper using fine forceps and lined up under a dissection microscope so that the anterior and posterior poles were all aligned in the same orientation. Excess moisture was soaked up with more filter paper and glass coverslips with thin strips of double-sided tape were gently pressed onto the embryos to transfer them. The attached embryos were left to desiccate for 30 s – 2 min, and the level of desiccation was carefully monitored under the dissection microscope because overly desiccated embryos will not develop, while insufficiently desiccated embryos may leak after injection. At the immediate onset of small wrinkles forming on the surfaces of the embryos, they were covered with a small blob of a 1:9 mixture of halocarbon oils 27 and 700 (Sigma-Aldrich, USA) to prevent further desiccation, and the coverslips were moved to a slide under a BA400 Motic light microscope (Motic, Hong Kong).

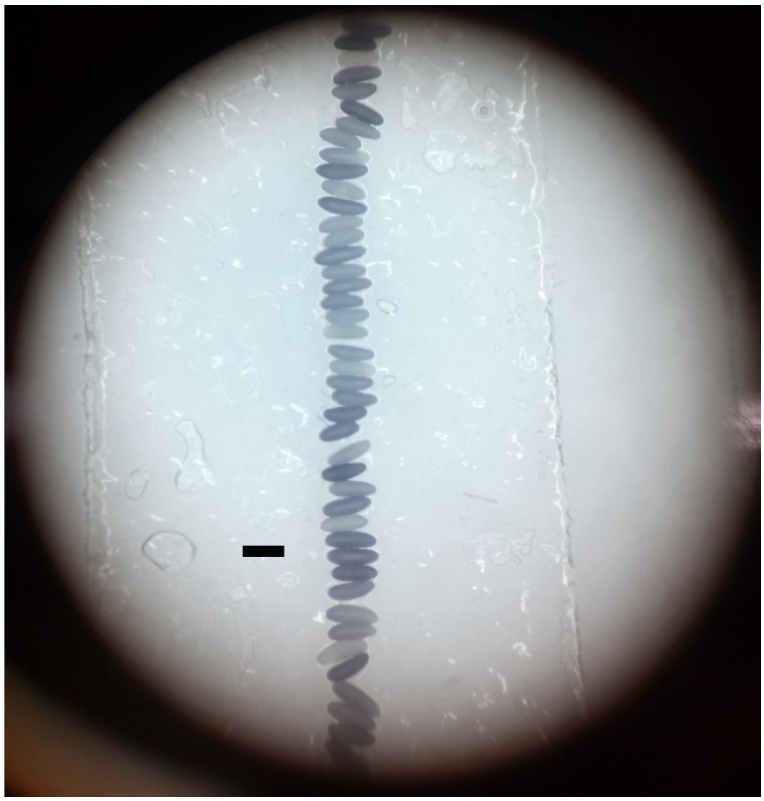


Figure 2.7 Ae. aegypti *embryos lined up with the posterior "pointy" poles facing right and attached to a microscope slide, showing various levels of maturity. Younger embryos are lighter grey while more mature eggs are darker grey due to the gradual melanisation of the chorion. Scale bar is 1 mm.*

Microinjection needles were fashioned from aluminosilicate glass filaments with a Sutter P-2000 needle puller (Sutter Instrument Company, UK) using a pre-existing program optimised at Oxitec for mosquito embryos with the following settings: HEAT: 420, FIL: 120, VEL: 50, DEL: 200, PUL: 140. The needles were loaded with ~2–4 µl of injection mix using a Microloader pipette tip (Eppendorf, Germany) and bevelled for approximately 10 s using an Intracel LTD bevelling machine (Sutter Instrument Company, UK). Bevelled needles were transferred to a MN-151 micromanipulator (Narishige, Japan) connected to a FemtoJet air compressor (Eppendorf, Germany) and moved into view under the BA400 microscope. When not in use, the needles were lowered into the halocarbon oil mixture described above to prevent the injection mix evaporating and blocking the needles.

2.3.3.2 Microinjection

The micromanipulator was used to position the needles at an angle of 10–25° and each embryo on the coverslips was injected in the posterior pole with a small amount (approximately 0.2 – 0.5 nl) of injection mix, so that a slight clearing of the yolks was visible. The injection pressure was adjusted with the FemtoJet according to the injection mix viscosity and level of desiccation of the embryos, and the back pressure was adjusted to prevent yolk flowing into the needles.

After injection, each coverslip was transferred to a metal rack in a hatch pot containing deionised water to allow the oil to run off. At the end of an injection day, the water and oil was drained, filter paper moistened with damp cotton wool was placed into the hatch pot, Parafilm was stretched over the top and the pot was kept in insectary conditions.

2.3.3.3 Screening for transgenic progeny

After 4 days, the matured eggs were hatched and G_0 larvae were reared to pupation using the same protocol described above, but in hatch pots rather than trays. In some cases, hatched L1 larvae were transferred to a new deli pot using a glass pipette to remove them from residues of halocarbon oil.

Pupae were sexed and placed in same sex pools in cages, and backcrossed to uninjected (wild type) individuals of the same strain. 1 or 2 G_0 males were crossed to 5 or 10 wild type females, respectively; while up to 10 G_0 females were crossed to up to 5 wild type males. The pooled cages were blood fed, eggs collected and hatched, and G_1 larvae were reared in deli pots as described.

L2 or L3 G_1 larvae were transferred to a weigh boat, immobilised on ice if required, and screened for fluorescence with UV light in a dark room using a Leica MZFLIII or an Olympus SZX12 fluorescence microscope. The presence of fluorescence, either blue from the *AmCyan* or red from the *DsRed2* marker genes, depending on the construct injected, was taken as evidence of successful transgenesis. Positive transformants were sorted and reared to pupation.

2.3.3.4 Establishment of transgenic lines

In cases where transgenic lines were needed, transformed pupae were placed in cages and reared over multiple generations according to standard protocols to generate inbred families. Given the low frequency of transformation in *Ae. aegypti*, positive G_1 transformants were assumed to be derived from a single transformation event in one of the G_0 individuals in the respective pooled cage. Thus, separate lines were named according to the pool from which the G_1 eggs were obtained; for instance, lines OX5226A and OX5226B could be expected to be descended from transformed G_0 individuals in pools A and B, respectively.

2.3.3.5 Reverse Transcription-PCR on transgenic embryos

Reverse Transcription-PCR (RT-PCR) was conducted on OX5226 and AWT (as a wild type control) embryos 0-4 h after oviposition to validate the expression of the integrated *cas9* gene. Embryos were collected from blood-fed females on a moistened Whatman filter paper and transferred to RNase-free tubes (Eppendorf, Germany). RNA was immediately extracted using the Total RNA Purification Plus Kit (Norgen Biotek Corp., Canada), including a gDNA column extraction step to remove all DNA, and the eluted RNA was quantified using a nanophotometer. 500 ng of each sample was converted to cDNA using the RevertAid RT Reverse Transcription Kit (ThermoFisher Scientific), along with a reaction containing no RevertAid reverse transcriptase as a no-RT control to ensure *cas9* was detected in mRNA rather than in any residual gDNA in the samples.

PCR reactions were run with BioTaq polymerase (PCR Biosystems) and primers targeting *cas9* (Appendix 2.2), using 1 µl of cDNA from each OX5226, WT and no-RT sample, water as a negative control, and OX5226 plasmid DNA as a positive control. Fragment sizes of the PCR products were examined using gel electrophoresis.

2.3.4 Design of injection experiments

The overall experimental design was divided into three sets of injections, summarised in Figure 2.8.

1. Initially, CRISPR-mediated integration was attempted at three putative male-specific genome regions by injecting LWT with CRISPR components including Cas9, along with each of the three template plasmids (OX5167-5169) (Figure 2.8, panel 1). The injections were carried out in two rounds, with increasing numbers of embryos.
2. Next, *piggyBac* insertion of the *nanos-cas9* plasmid (OX5226) was attempted by injecting the plasmid into LWT along with *piggyBac* helper mRNA (Figure 2.8, panel 2).
3. Finally, CRISPR-mediated knock-out and integration were attempted at the genome region surrounding the male-specific gene *Nix* by injecting embryos from the *piggyBac*-transformed lines and another lab strain, the inbred AWT Family 2, with CRISPR components:
    - Two sets of transgenic embryos, OX5226A and OX5226B, were injected only with sgRNAs with and without Cas9 enzyme (knock-out).
    - Two sets of transgenic embryos, OX5226A and OX5226B, were injected with all CRISPR components except Cas9, along with the template plasmid OX5346 (integration).
    - One set wild type embryos, AWT Family 2, were injected with all CRISPR components including Cas9, along with the template plasmid OX5346 (integration) (Figure 2.8, panel 3).

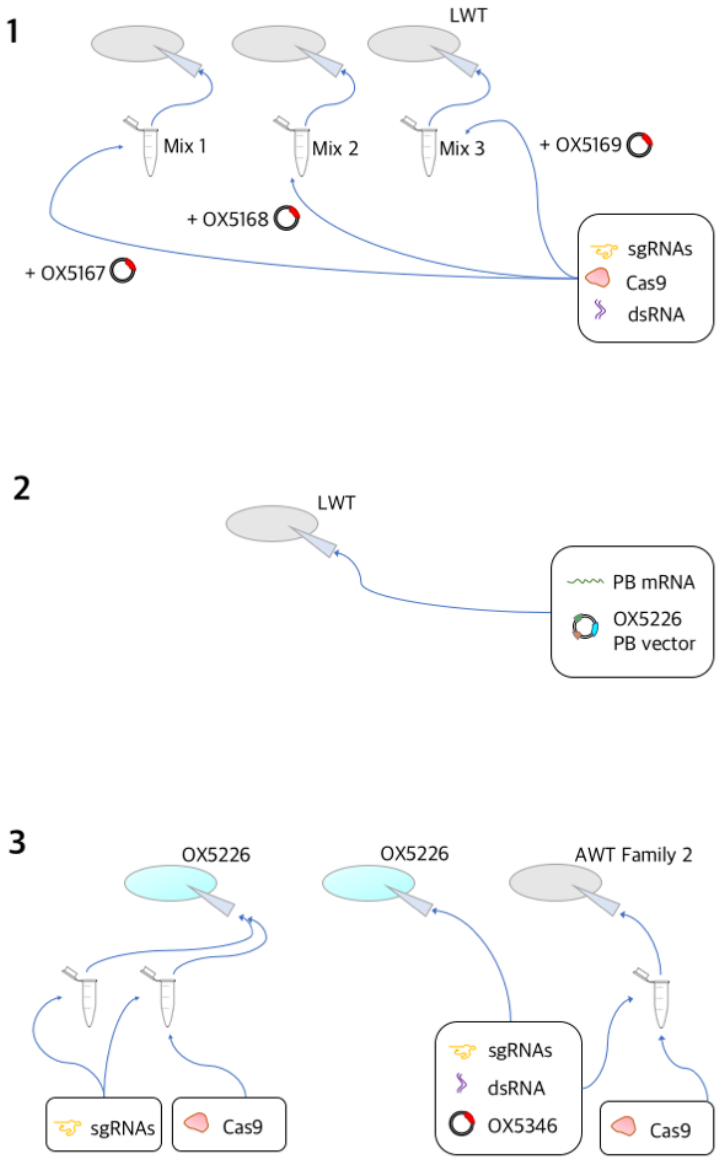


Figure 2.8 Design of the injection experiments. ***1*** *In the first CRISPR knock-in experiment, Latin wild type (LWT) embryos were injected with* ku70 *dsRNA, Cas9 enzyme, specific sgRNAs targeting one of three male-specific sequences, and a CRISPR integration DNA plasmid containing sequence flanking the targets cleaved by each set of sgRNAs along with the marker gene* DsRed *(OX5167-5169).* ***2*** *In the* piggyBac *transformation experiment, LWT embryos were injected with the germline Cas9 expression/*AmCyan *marker vector OX5226 and* piggyBac *helper mRNA.* ***3*** *In the subsequent CRISPR knock-in experiment, OX5226* cas9^{+} *embryos were injected with sgRNAs targeting the eye pigmentation gene* kmo *with and without Cas9 enzyme; and with* ku70 *dsRNA, sgRNAs targeting both* kmo *and the M locus gene* Nix, *and a CRISPR integration DNA plasmid containing* Nix *flanking sequence along with* DsRed *(OX5346). Embryos from the inbred strain Asian wild type Family 2 were also injected with these components along with Cas9 enzyme.*

2.4 Results

2.4.1 First CRISPR knock-in experiment: Targeting three putative male sequences

2.4.1.1 First round of injections

Table 2.2 shows the results for the first round of injections using the three M candidate targets.

Table 2.2 Results of Round 1 of the CRISPR integration experiment targeting three M locus candidates.

Mix name	Construct	Eggs injected	G_0 larvae hatched	G_0 % survival	G_1 transgenic larvae
CRISPR Mix 1	OX5167	1316	41	3.1	0
CRISPR Mix 2	OX5168	1224	24	2.0	0
CRISPR Mix 3	OX5169	1320	31	2.3	0
Total		3860	96	2.5	0

The survival was low and no transgenic G_1 larvae were detected for any of the constructs. Given that the efficiency of CRISPR-mediated knock-in is expected to be low, the number of injected eggs was increased to maximise the chances that transformation would occur.

2.4.1.2 Second round of injections

Table 2.3 Results of Round 2 of the CRISPR integration experiment targeting three M locus candidates.

Mix name	Construct	Eggs injected	G_0 larvae hatched	G_0 % survival	G_1 transgenic larvae
CRISPR Mix 1	OX5167	2496	29	1.2	0
CRISPR Mix 2	OX5168	2426	37	1.5	0
CRISPR Mix 3	OX5169	2495	19	0.8	0
Total		7417	85	1.1	0

Table 2.3 shows the results for the second round of injections were also negative. Even factoring in low survival and low efficiency of transformation, the absence of any transgenic mosquitoes after approximately 12,000 injections suggested that other factors might be interfering with the success of transgenesis. For instance, the presence of the Cas9 protein within the injection mix may have increased the toxicity of the injections. Additionally, despite the evidence that the sequences targeted are male-specific, nothing else is known about them; for instance, they may be transcriptionally inactive and consequently any successful integration would not be expressed.

To overcome these difficulties, two different approaches were taken: choosing new target sequences, and attempting to utilise germline Cas9.

2.4.2 Injection of endogenous germline Cas9 *piggyBac* vector

Previous work on other insects (mostly *D. melanogaster*) has shown that having Cas9 produced in the germline during early development can allow more efficient transgenesis as it removes the need to include the protein in the injection mix (Kondo and Ueda, 2013; Ren *et al.*, 2013; Gratz *et al.*, 2014). It was also achieved in medfly (*Ceratitis capitata*) at Oxitec Ltd. (R. Turkel, unpublished data). At the time that these experiments were carried out germline Cas9 expression had not been

demonstrated in *Ae. aegypti* in published literature, however it has subsequently been reported in Li *et al.* (2017).

10 pools of G_1 larvae were obtained from G_0 survivors injected with the *piggyBac* construct. Of these, one contained transgenic individuals (Table 2.4). These were separated by brightness of the fluorescent phenotype and treated as two separate lines, OX5226A and OX5226B, in case they resulted from separate integration events (Figure 2.9). The two lines were enriched but not made to be homozygous.

Table 2.4 Results of the piggyBac transformation experiment with the endogenous Cas9 construct. G_1 transgenic larvae refers to the total number of hatched progeny from backcrosses of G_0 survivors that displayed blue fluorescence.

Construct	Eggs injected	G_0 larvae hatched	G_0 % survival	G_1 transgenic larvae
OX5226	2184	46	2.1	49

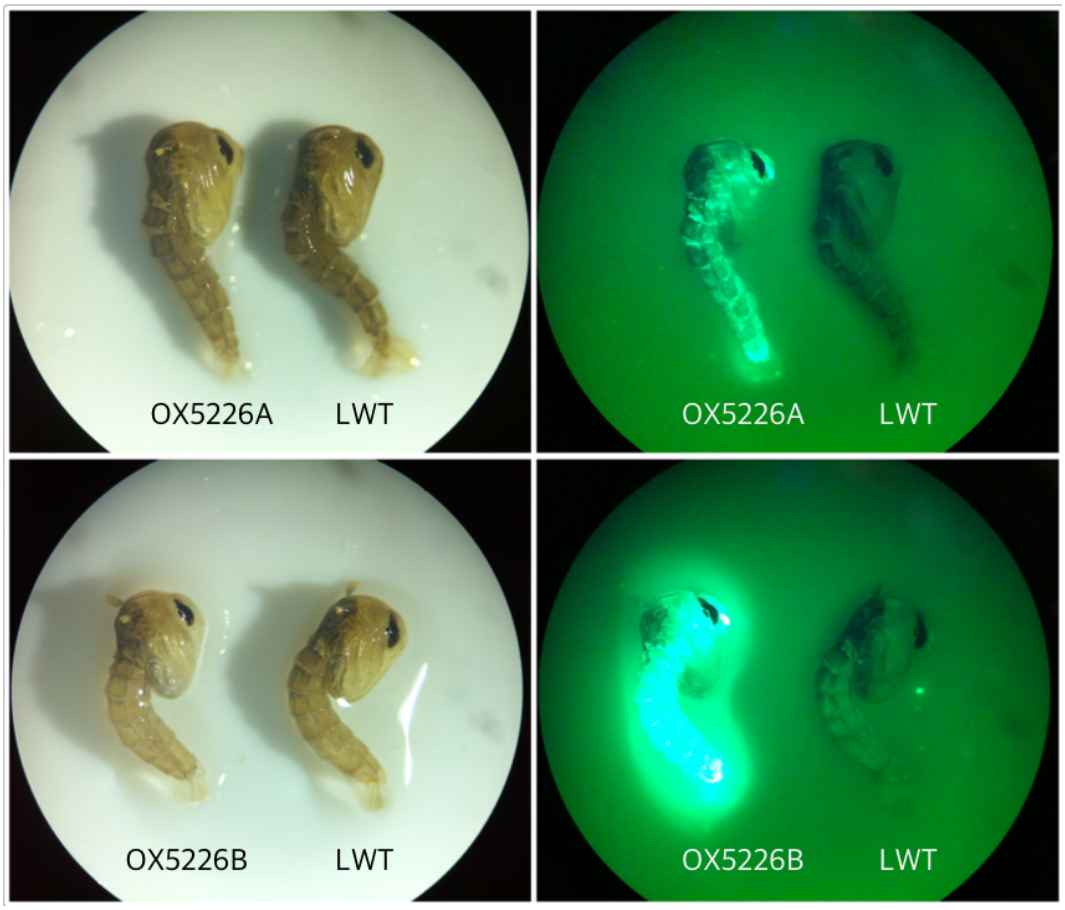


Figure 2.9 Latin wild type (LWT) and OX5226 Ae. aegypti *pupae expressing the AmCyan blue fluorescent protein, photographed under white light (left) and fluorescent light (right). Different intensities of fluorescence were visible, and assumed to be due to two different integration events and designated as two lines: OX5226A – dim (top) and OX5226B –bright (bottom).*

RT-PCR showed the presence of *cas9* mRNA in the embryos in both lines (Figure 2.10), indicating that they could be used to carry out further transformation using CRISPR/Cas9 and potentially result in higher transformation rates.

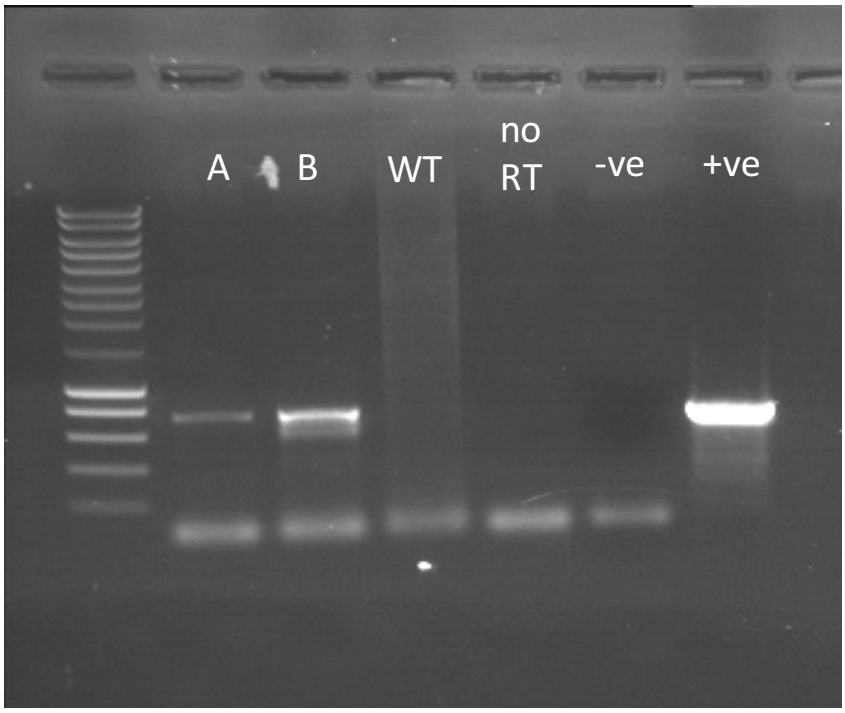


Figure 2.10 RT-PCR of endogenous germline Cas9 from Ae. aegypti *embryos. A and B show a positive result for* cas9 *in the early embryos of the two putative* cas9+ *OX5226 lines, whilst no* cas9 *cDNA was identified in the embryos of wild type (WT) embryos. The 'no RT' control contained no reverse transcriptase in the reverse transcription step. The negative control contained no template DNA and the positive control contained only the OX5226 plasmid DNA as the template in the PCR step.*

2.4.3 Second CRISPR knock-in experiment: Targeting the M locus gene *Nix* using germline expression of Cas9

During the time that the first CRISPR knock-in experiment was being conducted, an M locus gene called *Nix* was identified (Hall *et al.* 2015; more on the background of this gene and the further study that was conducted on it by the author is given in Chapter 3). This was used as an alternative M locus target to the three previous sequences into which to integrate DNA, due to the strong evidence that it is male-specific.

The design of the injection experiment is summarised in Figure 2.8, panel 3. The OX5346 construct, containing flanking sequence surrounding *Nix* derived from BAC library sequencing (Chapter 3) to act as a template for HDR (see Methods; 2.3.2.1),

was injected into OX5226 *cas9*$^{+}$ embryos, along with sgRNAs targeting *Nix* (see Methods; 2.3.2.4) and dsRNA to knock down *ku70*.

Although *Nix* is present in geographically varied *Ae. aegypti* strains (see Chapter 3.4), it is possible that the surrounding genome region is variable in different populations. The flanking sequence in the OX5346 plasmid is from the inbred AWT Family 2 strain, while the OX5226 strain was transformed from LWT; consequently, there may be sequence variation between LWT and AWT in the corresponding region adjacent to *Nix* that would reduce the effectiveness of the AWT Family 2/OX5346 sequence as a template for HDR. Therefore, to maximise the likelihood of successful CRISPR-mediated integration, embryos from AWT Family 2 – which would have DNA more closely matching the OX5346 flanking sequence – were also injected with the same components as the OX5226 *cas9*$^{+}$ embryos, with the addition of Cas9 enzyme.

Additionally, all injection mixes included sgRNAs targeting the eye pigmentation kynurenine 3-monoxygenase gene *kmo*, knockout of which results in a mutant white-eye phenotype (Aryan *et al.*, 2013; Basu *et al.*, 2015). These sgRNAs were included in the injection mixes for the knock-in experiments, and were also injected alongside sgRNAs targeting *Nix* into OX5226 without the OX5346 plasmid template or *ku70* dsRNA. Inspection for altered eye pigmentation and morphological feminisation due to somatic knockout of *kmo* and *Nix* could then be used as a separate test for NHEJ. Therefore, even with the NHEJ pathway suppressed with RNAi in the integration experiment, some mutations may occur in the mosquitoes injected with sgRNAs only, which could be used to validate the success of CRISPR gene editing.

Table 2.5 Results of the second CRISPR integration experiment targeting the M locus gene Nix *in the two germline Cas9 lines OX5226A and OX5226B, along with the wild type strain AWT Family 2.*

Strain injected	Construct	Eggs injected	G_0 larvae hatched	G_0 % survival	G_1 mutated
OX5226A	(sgRNAs only)	1555	14	0.9	0
OX5226B	(sgRNAs only)	1542	3	0.2	0
OX5226A	OX5346	1290	100	7.8	0
OX5226B	OX5346	1845	121	6.6	0
AWT Family 2	OX5346	2374	89	3.7	38
Total		8606	327	3.8	38

G_0 survivors from the injections of sgRNAs, dsRNA and the OX5346 construct into the *cas9*$^{+}$ OX5226 embryos were backcrossed to OX5226 in 18 pools, but no integration of OX5346 was detected in any of the G_1 progeny.

17 pools of G_0 survivors were obtained from injecting *cas9* AWT Family 2 embryos with sgRNAs, dsRNA, the OX5346 construct and Cas9 enzyme, of which one pool produced 38 transgenic G_1 progeny (Table 2.5): 19 males and 19 females expressed the fluorescent marker DsRed (Figure 2.11). Female transgenesis was unexpected as integration of the construct at *Nix* – which is only known to be present in the M locus – should result in only males exhibiting the transgenic phenotype, suggesting *DsRed* may have inserted off-site. David Navarro Paya screened the mutant progeny for *Nix* using PCR and found the gene was intact. PCR performed using primers for *DsRed* and the adjacent genomic flanking sequence in the construct resulted in a positive product in both males and females, suggesting that integration may have occurred at another location showing partial similarity to the target site (Supplementary Figure 1).

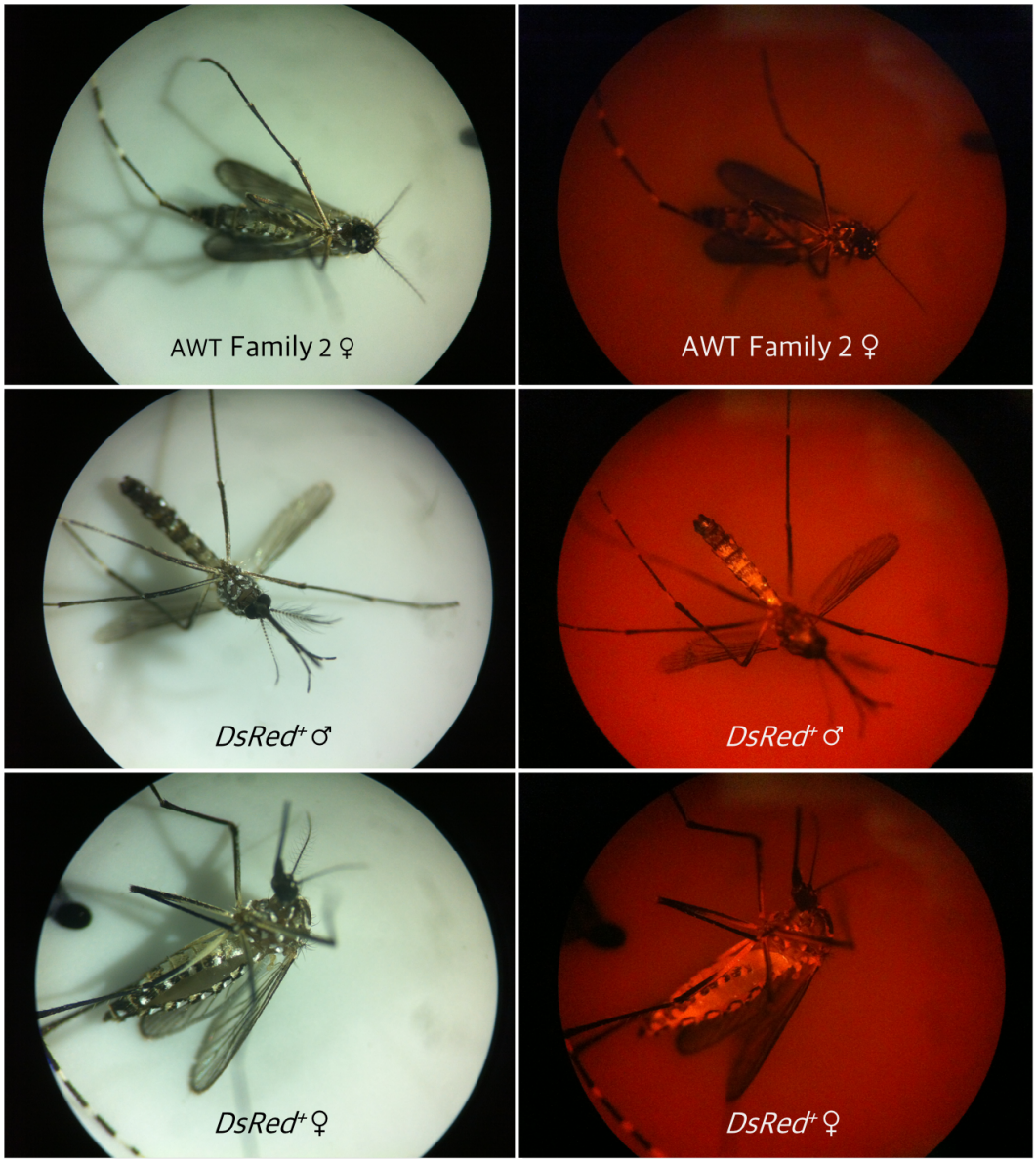


Figure 2.11 Asian wild type (AWT) Family 2 and transgenic DsRed+ Ae. aegypti G_1 adults, photographed under white light (left) and fluorescent light (right). 16 untransformed AWT Family 2 pools showed no fluorescence (top) while one pool showed red fluorescence in males (middle) and females (bottom), indicating the integration of the DsRed marker gene. The presence of fluorescence in females suggests the integration occurred at a site other than Nix, outside of the M locus.

The abdomens, antennae and eyes of adult G_0 mosquitoes from both injected strains, as well as survivors from the OX5226 embryos injected only with sgRNAs targeting *Nix* and *kmo*, were examined for evidence of somatic mutagenesis resulting from CRISPR gene editing. Some males displayed abnormal external genitalia: normal males possess a pair of gonocoxites on the terminal segments of the abdomen, which bear claw-like gonostyli, used to grip females during mating; yet some male G_0s had rotated or partially absent gonocoxites and gonostyli (Figure 2.12), similar to the *Nix* CRISPR knockout mutants observed by Hall *et al.* (2015), but no abnormalities were detected in the antennae, which retained the male "feathery" phenotype (Figure 2.13). No G_0s were detected with the presence of a somatic mosaic mutant white-eye phenotype (Figure 2.13), so pools of G_1s were inbred in the hope of detecting homozygous *kmo*$^-$/*kmo*$^-$ G_2 progeny exhibiting the full white-eye phenotype; however, none were detected.

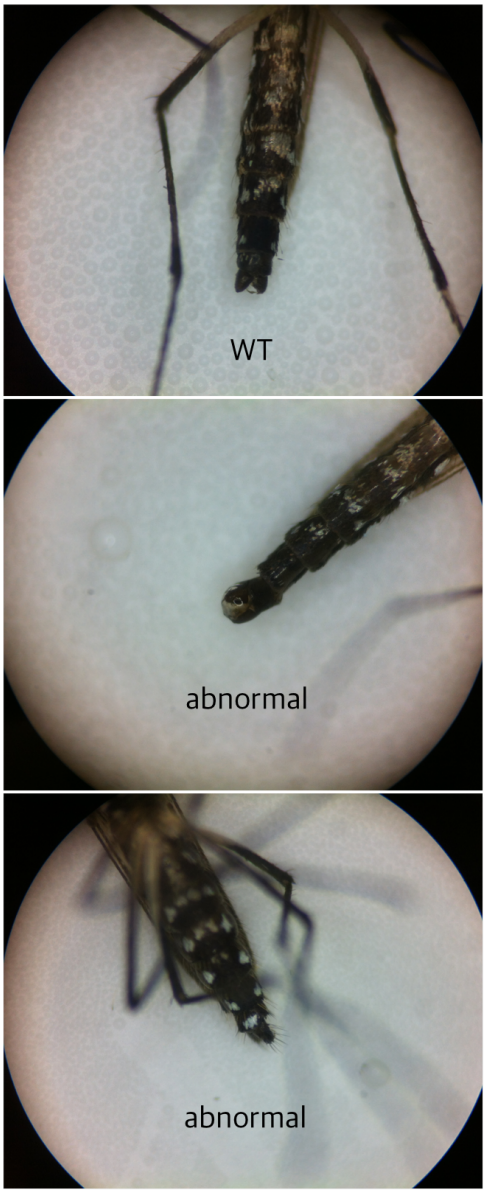


Figure 2.12 Male G_0 OX5226 Ae. aegypti *adults from embryos injected with sgRNAs targeting the M locus gene* Nix*, photographed under white light. Wild type (WT) males (top) had correctly formed genitals. Some males had abnormal genitals, such as rotation of the gonocoxites (middle) or absence of gonocoxites and/or gonostyli (bottom), suggesting possible CRISPR gene editing of* Nix *and consequent feminisation and/or disruption of male morphology.*

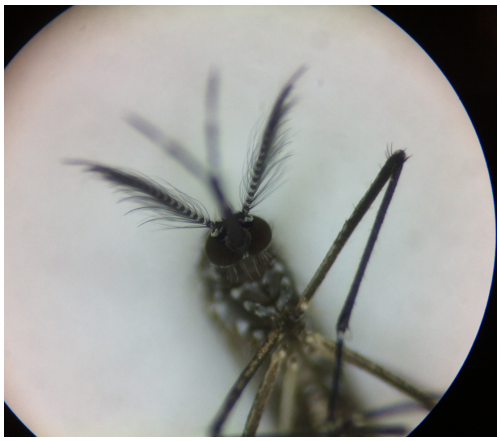


Figure 2.13 A G_0 OX5226 Ae. aegypti *adult from an embryo injected with sgRNAs targeting the eye pigmentation gene* kmo *and M locus gene* Nix, *photographed under white light. No somatic presence of the white-eye phenotype is present, indicating a lack of CRISPR gene editing. A typical wild type male "feathery" antenna phenotype can also be observed, further suggesting lack of feminisation from any CRISPR inactivation of* Nix.

2.5 Discussion

2.5.1 Establishment of an endogenous germline Cas9-expressing mosquito line

Gene editing with CRISPR/Cas9 has been successfully attempted in *Ae. aegypti* multiple times, utilising both NHEJ to perform targeted knock-outs and HDR to achieve integration of complex DNA templates. Previous studies have demonstrated it as a proof of concept and attempted to optimise its efficiency, while others have used it as a tool to investigate the function of particular genes (Basu *et al.*, 2015; Dong *et al.*, 2015; Hall *et al.*, 2015; Kistler *et al.*, 2015; Vinauger *et al.*, 2018). However, at the time that the experiments described in this chapter were being conducted, there were no studies in the published literature endeavouring to use germline expression of the Cas9 enzyme to improve mosquito control technologies, although it had been demonstrated in *Drosophila* and has since been applied in *Anopheles* (Gratz *et al.*, 2014; Gantz *et al.*, 2015; Galizi *et al.*, 2016; Hammond *et al.*, 2016). Subsequently, mutagenesis of Cas9-producing *Ae. aegypti* lines, transformed using *piggyBac* very similarly to the methods in this chapter, was demonstrated with high efficiency (Li *et al.*, 2017b). The germline Cas9 line generated in this chapter, OX5226, is therefore one of the first examples of this transgenic technique in *Aedes*.

There was strong evidence that Cas9 is present in the OX5226 germline: the AmCyan marker is expressed, indicating integration of the vector containing *cas9* (Figure 2.9); and *cas9* mRNA is detected in the early embryos (Figure 2.10). The survival of injected OX5226 embryos was slightly higher than previous rounds of injections (Table 2.5), implying the toxicity of the injection mix is lower. This suggests a potential improvement over using injected recombinant Cas9 protein, and OX5226 could be made to be homozygous and used for future CRISPR/Cas9 transformation attempts. Despite this, the lack of successful CRISPR editing means that more tests should be carried out to verify presence of functional Cas9 protein in the germline.

2.5.2 Poor efficiency of CRISPR mutagenesis

CRISPR-mediated integration via HDR was not achieved in wild type *Ae. aegypti* in the first CRISPR injection experiment, nor in the germline *cas9*$^{+}$ strain OX5226 in the subsequent experiment. In the second CRISPR injection experiment, a single knock-in mutation of *DsRed* was observed, however this did not occur at the desired M locus site targeted by the guide RNAs and did not result in male-specific transgenesis. Transgenesis arose by an unknown process; possible explanations include potential similarity of *Nix* flanking region to other parts of the genome resulting in HDR away from the target locus.

There was little evidence of CRISPR-mediated NHEJ. No inactivation of *kmo* due to CRISPR knockout could be observed in G_0 progeny or in subsequent generations after inbreeding as they all had normal eye pigmentation. The rotation and partial absence of external genitalia in injected males is very similar to the malformed and feminised phenotypes observed in previous experiments using CRISPR to knock out *Nix* (Hall *et al.*, 2015), suggesting successful CRISPR gene editing may have occurred, resulting in morphological feminisation. However, rotation of the posterior abdominal segments is observed in *Ae. aegypti* males just after eclosion, prior to correct orientation 1 – 2 days later (Clements, 1992). Therefore, this abnormal phenotype may simply have been observed in newly-emerged males, while the missing genitalia may have been due to damage during mating. The absence of feminised antennae supports the hypothesis that the abnormal genitalia resulted from causes other than CRISPR mutagenesis. Sequencing or performing high resolution melt analysis (HRMA) on *Nix* amplicons from the abdominal DNA of the injected males could have been used to check whether any indels were present.

It is unclear what lies behind the lower success of the experiments in this chapter compared to examples of effective CRISPR editing in *Aedes.* A recent study demonstrated efficient integration of two separate donor templates with homology arms of approximately 1 kb in various germline Cas9-producing lines, showing that there are no fundamental barriers to using a transgenic Cas9 system in this

mosquito species (Li *et al.*, 2017b). It may be that the amount of Cas9 in the embryos is too low to bind enough DNA for effective editing, or that not all OX5226 individuals produce Cas9 as the line was not made to be homozygous. However, maternal deposition of Cas9 can result in mutagenesis even in progeny that do not contain the transgene themselves (Lin and Potter, 2016), so this is unlikely to explain the low efficiency.

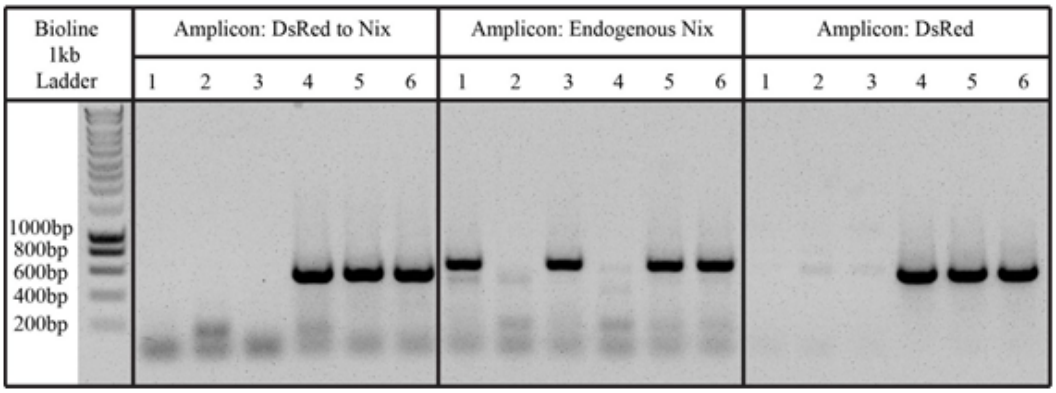
In any event, co-injection of recombinant Cas9 protein into wild type embryos was little more successful than using the *cas9*$^{+}$ embryos, as the one positive HDR event was not correctly targeted, indicating that the effectiveness was not dependent on the source of the enzyme. The off-target integration could have been due to the heterochromatic nature of the M locus, whereby fragments of repetitive sequence elsewhere in the genome may have formed the site of HDR. However, precise CRISPR-mediated integration has been demonstrated in the similarly heterochromatic *D. melanogaster* Y chromosome (Buchman and Akbari, 2018), suggesting that such issues should not completely bar the modification of the M locus.

2.5.3 Future directions

Overall, the success of CRISPR/Cas9 for the purposes of introducing functional constructs at targeted parts of the *Ae. aegypti* genome was very low. Male-specific transgenesis was not achieved, and only one instance of HDR was detected, in which the donor sequence was inserted at an undetermined site other than the one targeted. Use of an endogenous Cas9-producing line did not improve the outcome. Although this seems to suggest that the prospects of applying this technology to mosquito control are limited, the established effectiveness of CRISPR/Cas9 in *Aedes*, in addition to the positive results of incipient gene drive techniques in other species such as *Anopheles*, indicates that further research is worthwhile. Future work could draw from the unsuccessful avenues pursued in the experiments in this chapter to advance sex-specificity of genetic vector control strategies.

The results of these experiments also show that CRISPR-mediated knock-ins are an inefficient way of investigating the content of the *Ae. aegypti* M locus. Although CRISPR/Cas9 has been effectively utilised in many species to investigate gene functions, it did not prove to be useful for ascertaining male bias of genome fragments. Notwithstanding this, better understanding of the structure of the M locus is likely to enable improved mosquito control, whether using CRISPR or other techniques. The following chapters deal with alternative approaches for identifying male-specific fragments and sequencing the M locus.

2.6 Supplementary data



Supplementary Figure 1 PCR results of six mosquito samples with three sets of primers. The samples are ***1*** *male OX5226;* ***2*** *female AWT;* ***3*** *male AWT;* ***4*** *female* DsRed^{+} *mutant;* ***5*** *male* DsRed^{+} *mutant;* ***6*** *male* DsRed^{+} *mutant (second replicate). The first set of primers (left) targeted the intersection of* DsRed *and the homology arm containing the* Nix *sequence; the second set (middle) targeted* Nix *endogenously, rather than in the OX5346 donor; and the third set targeted the* DsRed *exon. The results show that* Nix *remains intact in all males, while both the female and male mutant progeny have the* DsRed *marker gene integrated at the same non-sex-specific site. Figure by D. Navarro Paya.*

Chapter 3 The sequence of a male-specific genome region containing the sex determination switch in *Aedes aegypti*

Part of this chapter has been published as

Turner, J., Krishna, R., Van't Hof, A.E., Sutton, E.R., Matzen, K. and Darby, A.C. (2018). The sequence of a male-specific genome region containing the sex determination switch in *Aedes aegypti*. *Parasites & Vectors*. **11**:549.
(See Appendix 1)

An earlier version of this manuscript was deposited on the preprint server *bioRxiv* and can be found at
https://doi.org/10.1101/122804

3.1 Abstract

Aedes aegypti is the principal vector of several important arboviruses. Among the methods of vector control to limit transmission of disease are genetic strategies that involve the release of sterile or genetically modified non-biting males, which has generated interest in manipulating mosquito sex ratios. Sex determination in *Ae. aegypti* is controlled by a non-recombining Y chromosome-like region called the M locus, yet characterisation of this locus has been thwarted by the repetitive nature of the genome. In 2015, an M locus gene named *Nix* was identified that displays the qualities of a sex determination switch. With the use of a whole-genome bacterial artificial chromosome (BAC) library, we amplified and sequenced a ~200kb region containing the male-determining gene *Nix*. In this study, we show that *Nix* is comprised of two exons separated by a 99kb intron primarily composed of repetitive DNA, especially transposable elements. *Nix* is an unusually large and highly repetitive gene, and exhibits features in common with Y chromosome genes in other organisms. We speculate that the lack of recombination at the M locus has allowed the expansion of repeats in a manner characteristic of a sex-limited chromosome, in accordance with proposed models of sex chromosome evolution in insects.

3.2 Introduction

At least 2.5 billion people live in areas where they are at risk of dengue transmission from mosquitoes, principally *Ae. aegypti*, with an estimated 390 million infections per year (Laughlin *et al.*, 2012; Bhatt *et al.*, 2013). Recently, the emergence of chikungunya and Zika viruses further highlights the public health importance of *Ae. aegypti* (Musso *et al.*, 2015; Fauci and Morens, 2016). Future mosquito control strategies may incorporate genetic techniques such as the sustained release of sterile or transgenic "self-limiting" mosquitoes (Alphey, 2014; World Health Organization, 2016). Given that only female mosquitoes bite and spread disease, there has been substantial interest in manipulating mosquito sex determination using these genetic techniques and others, including gene drive (Gilles *et al.*, 2014; Hoang *et al.*, 2016). Therefore, elucidating the genetic basis for sex determination could, for instance, facilitate production of male-only cohorts for release, or allow transformation of mosquitoes with sex-specific "self-limiting" gene cassettes.

Sex determination in insects is variable, and generally not well understood outside of model species (Charlesworth and Mank, 2010). Unlike the malaria mosquito *Anopheles gambiae* and *Drosophila* species, *Ae. aegypti* does not have heteromorphic (XY) sex chromosomes (Craig *et al.*, 1960). Instead, the male phenotype is determined by a non-recombining M locus on one copy of autosome 1 (Newton *et al.*, 1978; Clements, 1992; Toups and Hahn, 2010). This locus is poorly characterised because its highly repetitive nature has confounded attempts to study it based on the existing genome assembly (Hall *et al.*, 2015). The initial 1,376 Mb *Ae. aegypti* reference genome was assembled from Sanger sequencing reads in 2007 (Nene *et al.*, 2007), which are commonly not long enough to span the repetitive transposable elements that comprise a large proportion of the genome (Koren and Phillippy, 2015), and consequently the assembly was relatively low quality (Severson and Behura, 2012). Furthermore, the fact that both male and female genomic DNA was used for genome sequencing reduces the expected coverage of the M locus to one

quarter of the autosome 1 sequences, further obscuring candidate M locus sequences (Hall *et al.*, 2014).

Recently, a team of researchers was nevertheless able to identify *Nix*, a gene with male-specific, early embryonic expression. Knockout of *Nix* using CRISPR/Cas9 results in morphological feminisation of male mosquitoes along with feminisation of gene expression and female splice forms of the conserved sex-regulating genes *doublesex* (*dsx*) and *fruitless* (*fru*), strongly indicating that *Nix* is the upstream regulator of sexual differentiation (Hall *et al.*, 2015). The translated *Nix* protein contains two RNA recognition motifs and is hypothesised to be a splicing factor, acting either directly on *dsx* and *fru* or on currently unknown intermediates (Adelman and Tu, 2016; Figure 3.1). A comparison of sexually dimorphic gene expression in different mosquito tissue types also detected male-specific transcripts of *Nix* (Matthews *et al.*, 2016). An ortholog of *Nix* is present in *Ae. albopictus*, but it is not known if the two are functionally homologous (Chen *et al.*, 2015).

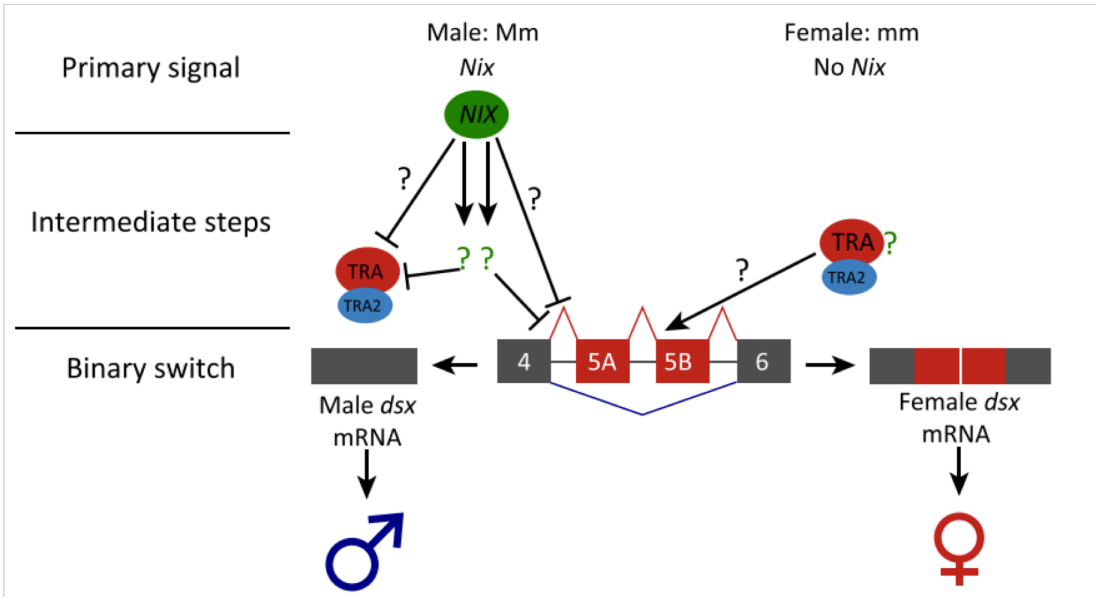


Figure 3.1 Schematic of a proposed sex determination cascade leading to the alternative splicing of doublesex *in* Ae. aegypti. *In males with Mm karyotype, the protein product of* Nix *(green oval) could interact with a Transformer/Transformer 2 complex (red and blue ovals), or unknown intermediate gene products (green question marks), to inhibit the female splice sites of* dsx. *The male splice form of* dsx *is produced, initiating downstream male development. Figure from Adelman and Tu (2016).*

To date, *Nix* has only been characterised as an mRNA transcript. To fully understand this gene's role in sex determination and to utilise this knowledge for vector control, it is essential to decipher its genomic context. For this purpose, this study identifies and describes the region of the M locus in which *Nix* is located.

3.3 Materials and methods

3.3.1 BAC library construction

A BAC library was constructed using living DH10b phage resistant *Escherichia coli* transfected with the pCC1BAC low copy number vector and *Ae. aegypti* genomic DNA from a DNA pool of approximately 50 sibling males (Amplicon Express, USA). Average insert size was 130 kb and the estimated coverage was ~5× for autosomal regions (~2.5× for sex specific regions). The male siblings were from one family (known as Family 2) of the Asian wild type (AWT; also known as My1) laboratory strain originating in Jinjang, Kuala Lumpur, Malaysia in the 1960s (described in Chapter 2.3.1.1 and in Lacroix *et al.* 2012), after five generations of full-sib mating. The final BAC library comprised 73,728 clones contained in 195 384-well plates. The plates were stored at -80°C.

3.3.2 BAC library screening

3.3.2.1 Outline of superpooling and matrixpooling method

96-well plates of superpools and matrixpools were supplied to allow screening of the BAC library for sequences of interest using two rounds of PCR, enabling the particular BAC clone or clones containing the sequence to be determined, as described in previous studies (Tao *et al.*, 2002; Bouzidi *et al.*, 2006). In the first round, DNA from each well in the superpool plate was used as a template for an individual PCR reaction using primers for the sequence of interest. The superpool plate comprised 28 superpools, containing purified DNA extracted and combined from 2,688 separately grown BAC clones, corresponding to a block of seven 384-well plates in the BAC library. Examining the results of the PCR reactions for positive products using gel electrophoresis revealed which superpools contained the clone of interest. Each superpool had a corresponding matrixpool plate on a section of a 96-well plate.

In the second round of PCR, DNA from each well in the matrixpool plates corresponding to the superpools of interest was used in PCR reactions, using the same target primers as in the first round. Similarly to the superpools, each well of the matrixpools contained purified DNA from a number of individual BAC clones, combined in a unique manner such that the locations of two positive reactions for each of the plate, row and column matrices could be used to decipher the coordinates of the clone of interest within the BAC library (Figure 3.2).
PCR reactions were run in 20 µl volumes with LongAmp Taq polymerase (New England Biolabs, USA), using the routine protocol for that enzyme but without a final 10 minute extension step. Superpools and matrixpools were stored at -20°C when not in use.

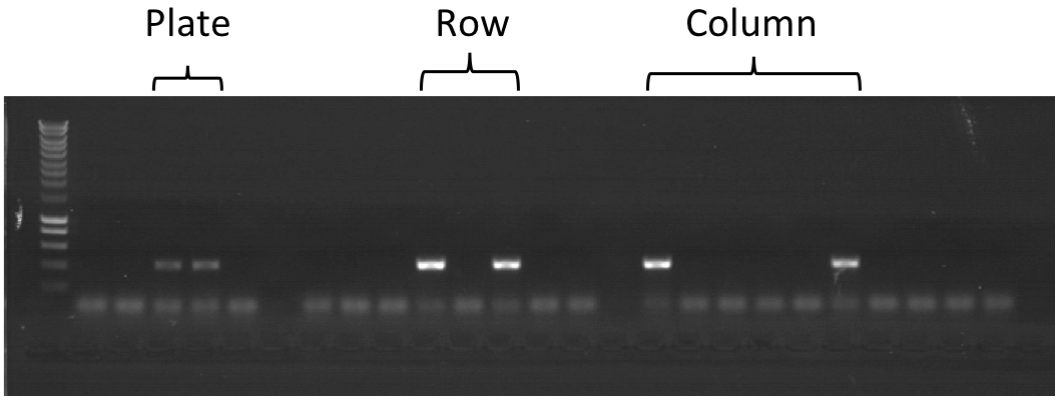


Figure 3.2 An example of PCR screening of a matrixpool with primers targeting a sequence of interest. Amplified products exist for wells 3 and 4 of the plate matrix, 4 and 6 of the row matrix, and 1 and 6 of the column matrix. With the use of a key, the coordinates of the BAC clone containing the sequence can be determined in the 7-plate superpool block – and therefore in the BAC library as a whole.

3.3.2.2 Screening for Nix

The BAC library was PCR screened using two sets of primers (Nix1F 3'-TTGAGTCTGAAAAGTCTATGCAA-5', Nix1R 3'-TCGCTCTTCCGTGGCATTTGA-5', Nix2F 3'-ACGTAGTCGGCAACTCGAAG-5', Nix2R 3'-CTGGGACAAATCGAACGGAA-5') based on the complete coding sequence of *Nix* (GenBank accession number KF732822). The first primer set was

also used to screen for *Nix* in the genomic DNA of six male and six female individuals each from two wildtype *Ae. aegypti* strains. Screening of the library resulted in four positive clones – two for each primer pair.

3.3.2.3 Screening for additional M locus sequences

After the BACs containing *Nix* were sequenced and assembled, the BAC library was PCR screened using primers targeting a number of other candidate male-specific sequences, in order to locate additional BACs containing M locus sequences and extend the sequenced region. These candidates were either based on the assembled *Nix* region or on other sequences thought to be within or linked to the M locus. These were:

- The 3' and 5' end sequences of the assembled *Nix* BACs.
- A region of predicted coding sequences (CDSs) in the assembled *Nix* BACs that show strong expression.
- The gene *myo-sex*, known to be tightly linked to the M locus (Hall *et al.*, 2014).
- A BAC clone (NDL62N23) originally from the Notre Dame Liverpool (NDL) *Ae. aegypti* BAC library (Jiménez *et al.*, 2004), which was found to be male-biased and subsequently sequenced (Hall *et al.*, 2014).
- The 18S ribosomal DNA (rDNA) locus, known to be close to the location of the M locus (Timoshevskiy *et al.*, 2013; Hall *et al.*, 2014; Timoshevskiy *et al.*, 2014).
- Seven candidate male-biased contigs from the AaegL3 assembly identified using the differential male-female coverage pipeline, three of which were used as targets for CRISPR knock-in, as described in Chapter 2.3.2.1: AAGE02035037.1(1-6260), AAGE02035965.1(1-4650), AAGE02035016.1(1-6296), AAGE02035557.1(1-5425), AAGE02036067.1(1-4250), AAGE02035994.1(1-4545), and AAGE02034767.1(1-6813).

For sequences that were identified in the BAC library, the primers were also used to screen for the sequences in the genomic DNA of six male and six female AWT individuals. Primers were designed using the NCBI primer designing tool (ncbi.nlm.nih.gov/tools/primer-blast), which uses Primer-BLAST (Ye *et al.*, 2012).

3.3.3 Isolation and sequencing of BAC clones

BAC clones were plated onto selective chloramphenicol agar plates and incubated overnight at 37°C. Single colonies were picked and dipped into flasks of 500 ml LB broth containing 12.5 µg/ml chloramphenicol, and flasks were incubated for 16 h in a shaking incubator at 37°C, 250 rpm.

BAC DNA was extracted from the bacterial cultures using the Qiagen Plasmid Maxi Kit (Qiagen, Germany), run on a 0.5% agarose gel at 30V for 16 h to confirm large fragment size, and quantified using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific). DNA from the four *Nix*-containing BACs was pooled before SMRTbell library preparation (PacBio, USA), and sequenced on a single SMRTcell using P6-C3 chemistry on the PacBio RS II platform (PacBio, USA).

3.3.4 Data analysis

The sequence data was trimmed to remove vector sequences and adaptors prior to assembly with the CANU version 1 assembler (Berlin *et al.*, 2015), followed by sequence polishing with QUIVER (Chin *et al.*, 2013).

BLASTN (Altschul *et al.*, 1990) was used to assess the uniqueness of the assembled *Nix* region compared to the *Aedes aegypti* Liverpool reference genome AaegL3 and the newer Aag2 cell line assembly. Illumina data generated from male and female genomic DNA (accession numbers SRR871496–SRR871497 and SRR871499–SRR871500) and RNA (accession numbers SRR1585314- SRR1585319; Appendix 2.1) were mapped to a combined reference containing the assembled *Nix* region added to the AaegL3 genome. DNA samples were mapped with BOWTIE 2.2.1 (Langmead and Salzberg, 2012) using default parameters with –I 200 and -X 500,

and RNA-Seq data with TopHat 2.1.1 (Kim *et al.*, 2013) using default parameters. RNA-Seq data was processed using the Cufflinks 2.2.1 pipeline (Trapnell *et al.*, 2012; Trapnell *et al.*, 2013) to look for potential genes and male/female specific expression from the region.

Genes were predicted using Augustus (Keller *et al.*, 2011) and the *Ae. aegypti* model (Nene *et al.*, 2007), and repetitive regions described using RepeatMasker 4.0.6 (Smit *et al.*) and the *Ae. aegypti* repeat database.

After the release of the AaegL5 reference genome, the assembled *Nix* region was aligned to the corresponding region of the reference using MUMmer 4.0.0 (Kurtz *et al.*, 2004) to look for regions of similarity.

3.4 Results

3.4.1 The complete gene sequence of *Nix*

Four BAC clones positive for *Nix* (23M14, 42I18, 113D8, 127D7) assembled into a single region of 207 kb with no gaps and a GC content of 40.2% (submitted to the NCBI as accession KY849907). The presence of the *Nix* gene in the assembled BACs was confirmed by BLASTN. The whole gene was present in tiled BACs, though not completely within individual BAC clones. Neither *Nix* nor the complete region could be found in the AaegL3 or Aag2 reference genome assemblies. While *Nix* was originally identified in the genome-sequenced Liverpool strain (Hall *et al.*, 2015), PCR revealed that it is exclusively present in male genomic DNA from other geographically varied *Ae. aegypti* populations (Figure 3.3), further strengthening the evidence that it is wholly present in the M locus.

The newly released AaegL5 male assembly contains *Nix* (Matthews *et al.*, 2018), and the assembled BACs aligned to the corresponding region in AaegL5 with >99.9% identity, spanning a 2899 bp gap in the AaegL5 genome that is comprised mainly of repeats (Figure 3.4; Figure 3.6).

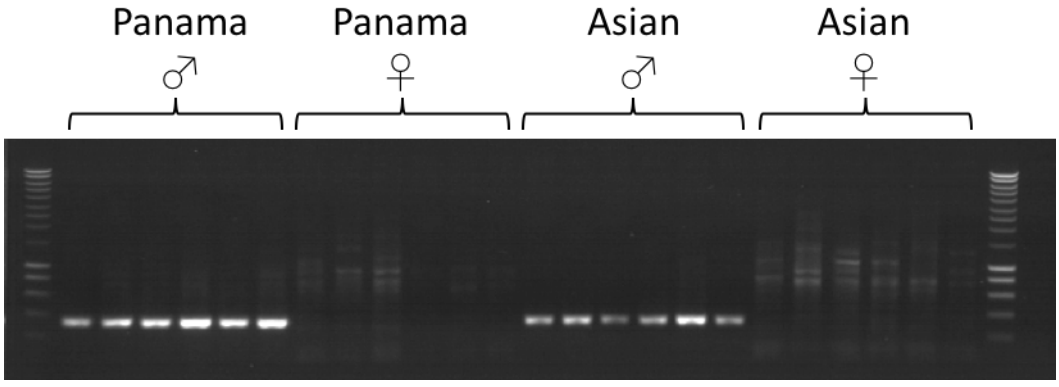


Figure 3.3 PCR screening of the M locus gene Nix *(exon 1) in 6 male and 6 female DNA of wild type* Ae. aegypti *strains.*

The *Nix* gene was found to be made up of two exons with a single intron of 99 kb (Figure 3.4). Although large introns are not uncommon in *Ae. aegypti* (average

intron length ~5000 bp) (Nene *et al.*, 2007), this intron is at the extreme end of intron sizes observed (Figure 3.5), especially considering the small size of its protein coding regions (<1000 bp). The gene structure is confirmed by Illumina RNA-Seq data clearly showing reads spanning the intron between the two exons (Figure 3.4). RepeatMasker identified approximately 55% of the sequenced region as repetitive, and the intron region of *Nix* as 72% repetitive (Table 3.1).

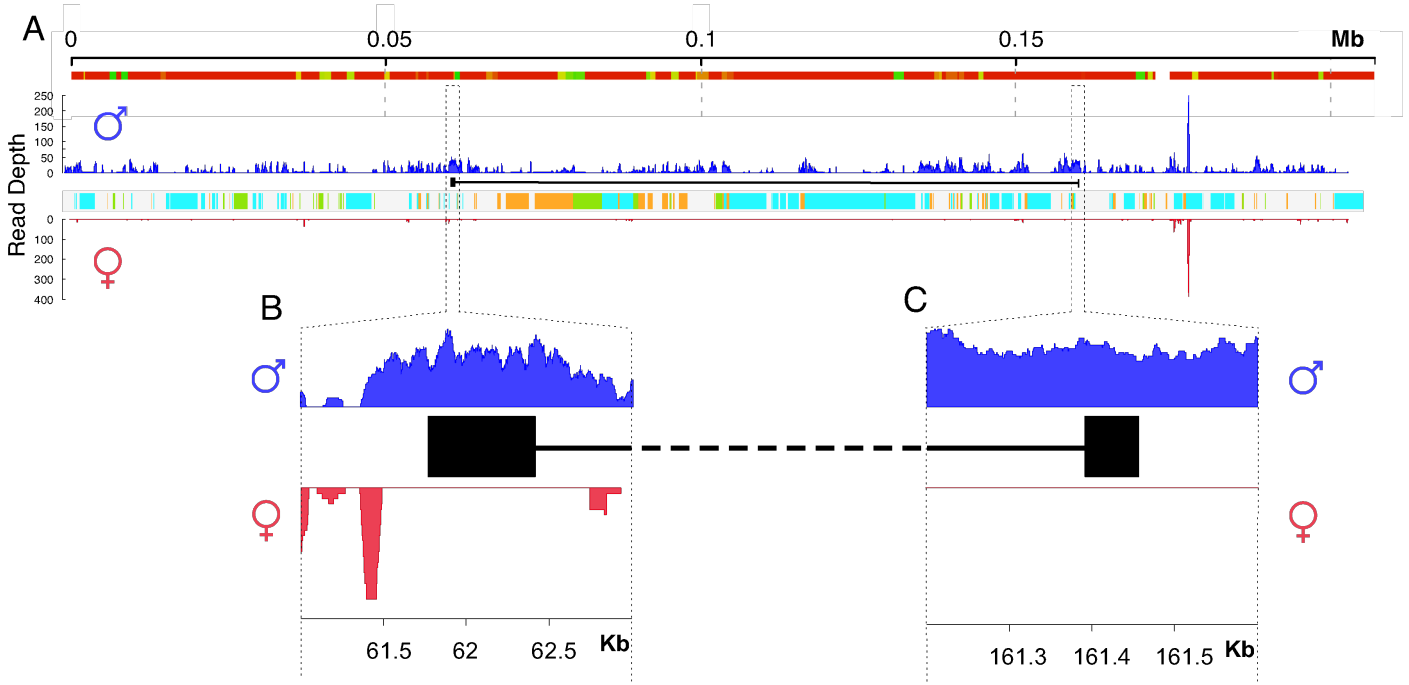


Figure 3.4 Structure and gene expression of the ~207 kb genomic region containing the Nix *gene.* Nix *is shown as two black boxes representing the exons, joined by a black line representing the intron. The top track of* ***A*** *shows the alignment of the sequence to the corresponding region of the reverse complement of the AaegL5 reference genome assembly, with colours representing percentage similarity (red: 100%; orange: >90%; green: >80%). Colours on the central track of* ***A*** *represent the classes of repetitive elements (orange: DNA transposons; cyan: Gypsy LTRs; green: Ty1/Copia LTRs). Blue histograms represent the coverage of RNA-Seq reads from adult male samples on the y axis; red histograms represent the coverage from adult female samples.* ***B*** *and* ***C*** *show enlargements of the first and second exons of* Nix *in the dotted regions in* ***A****, respectively.*

Table 3.1 Types and abundance of repeats in the 207kb assembled M locus region and 99kb Nix *intron, identified by RepeatMasker using the* Aedes aegypti *repeat library.*

	ENTIRE REGION		*NIX* INTRON REGION	
REPEAT TYPE	No. elements	% of sequence	No. elements	% of sequence
RETROELEMENTS	105	42.1%	49	51.0%
SINES	8	0.81%	5	1.11%
PENELOPE	3	0.08%	2	0.20%
LINES	24	5.43%	6	6.85%
L2/CR1/REX	4	0.13%	0	0%
R1/L0A/JOCKEY	13	3.87%	3	6.60%
RTE/BOV-B	3	1.33%	0	0%
L1/CIN4	1	0.02%	1	0.05%
LTR ELEMENTS	73	35.8%	38	43.0%
BEL/PAO	9	0.71%	3	0.87%
TY1/COPIA	16	11.3%	14	19.2%
GYPSY/DIRS1	48	23.8%	21	23.0%
DNA TRANSPOSONS	97	11.7%	69	20.1%
TC1-IS630-POGO	11	3.87%	11	9.04%
OTHER (MIRAGE, P-ELEMENT, TRANSIB)	1	0.06%	0	0%
UNCLASSIFIED	6	0.48%	3	0.22%
SMALL RNA	8	0.81%	5	1.11%
SATELLITES	1	0.75%	0	0%
SIMPLE REPEATS	19	0.34%	7	0.24%
LOW COMPLEXITY	3	0.07%	1	0.04%
TOTAL REPEATS		55.4%		71.6%

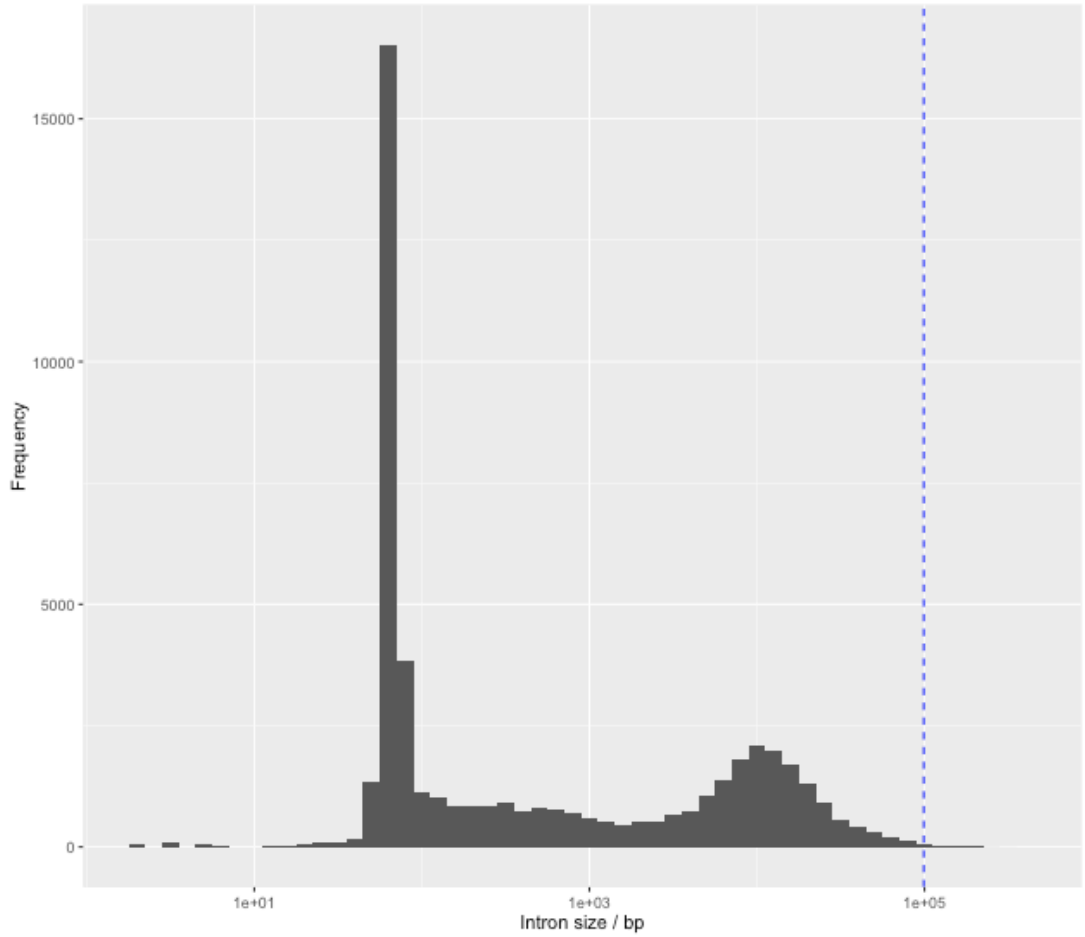


Figure 3.5 Intron size distribution in Aedes aegypti *Liverpool reference genome AaegL3. Blue dashed line indicates the size of the* Nix *intron relative other introns.* x *axis is transformed by log10.*

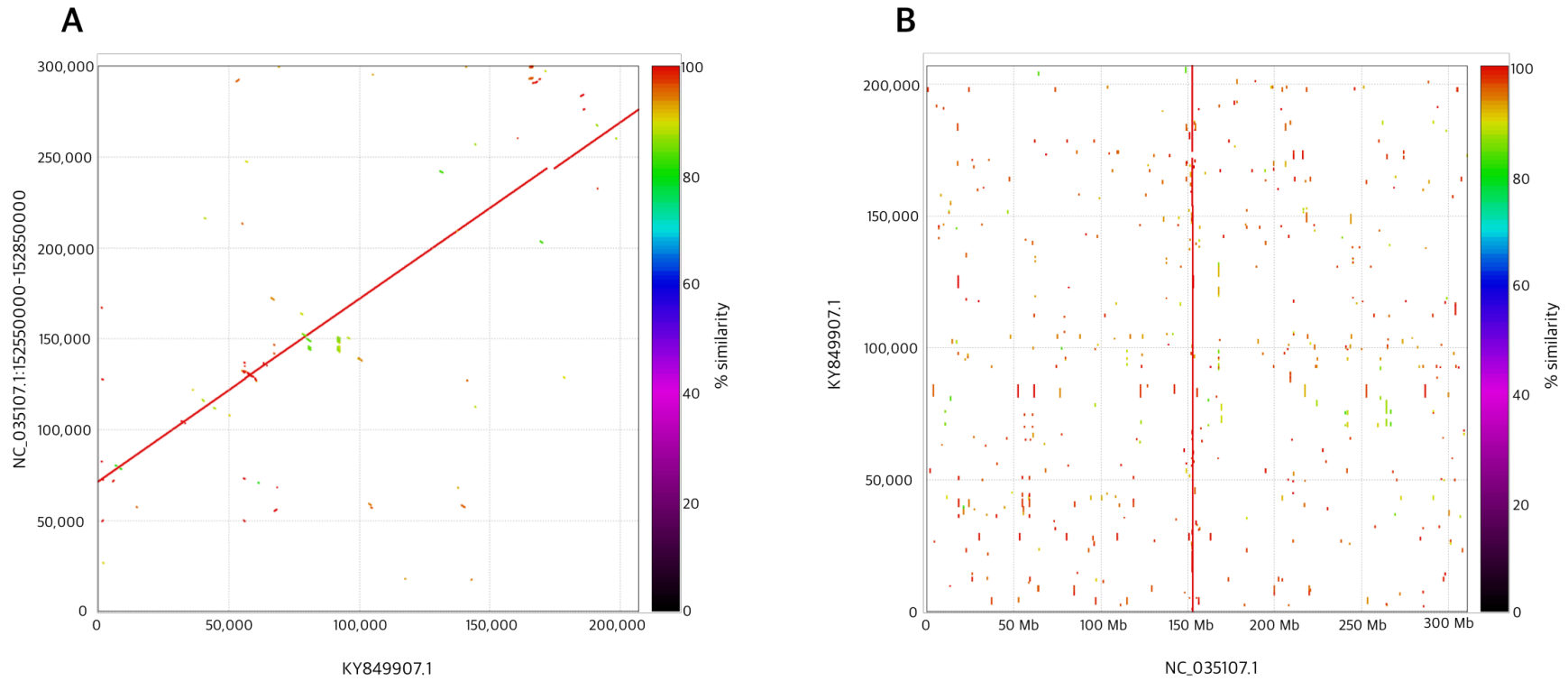


Figure 3.6 Alignment of the assembled 207 kb BAC region to chromosome 1 of the AaegL5 male reference assembly. ***A*** *The assembled* Nix *BAC sequence (KY849907.1;* x *axis) queried by the reverse complement of the corresponding region of the reference genome (NC_035107.1:152550000-152850000, indicating coordinates 152,550,000-152,850,000 on chromosome 1;* y *axis).* ***B*** *The assembled* Nix *BAC sequence (KY849907.1;* y *axis) queried against the reverse complement of chromosome 1 of the reference genome (NC_035107.1;* x *axis). Alignments and plots generated using MUMmer 4.0.0 (Kurtz* et al., *2004).*

3.4.2 Identification of additional male-biased BACs

To expand the sequenced region, the BAC library was screened for the 3' and 5' ends of the region, with the aim of identifying BACs containing adjacent genome sequences and "walking out" from *Nix* to encompass more of the M locus. Two primer pairs identified positive hits for BAC clones, but were not male specific in the wild type strain (Figure 3.7; Table 3.2).

In addition, an array of predicted CDSs and introns proximal to *Nix* (~170–175 kb) showed high depth of coverage of RNA-Seq reads (Figure 3.4), suggesting a potential functional role. However, BAC screening either returned no positive hits, or hits in many superpools, indicating that it is not a single-copy gene and is likely to be a transposable element with high expression. Positive hits were found in both male and female wild type gDNA, consistent with the high male and female expression levels shown in Figure 3.4.

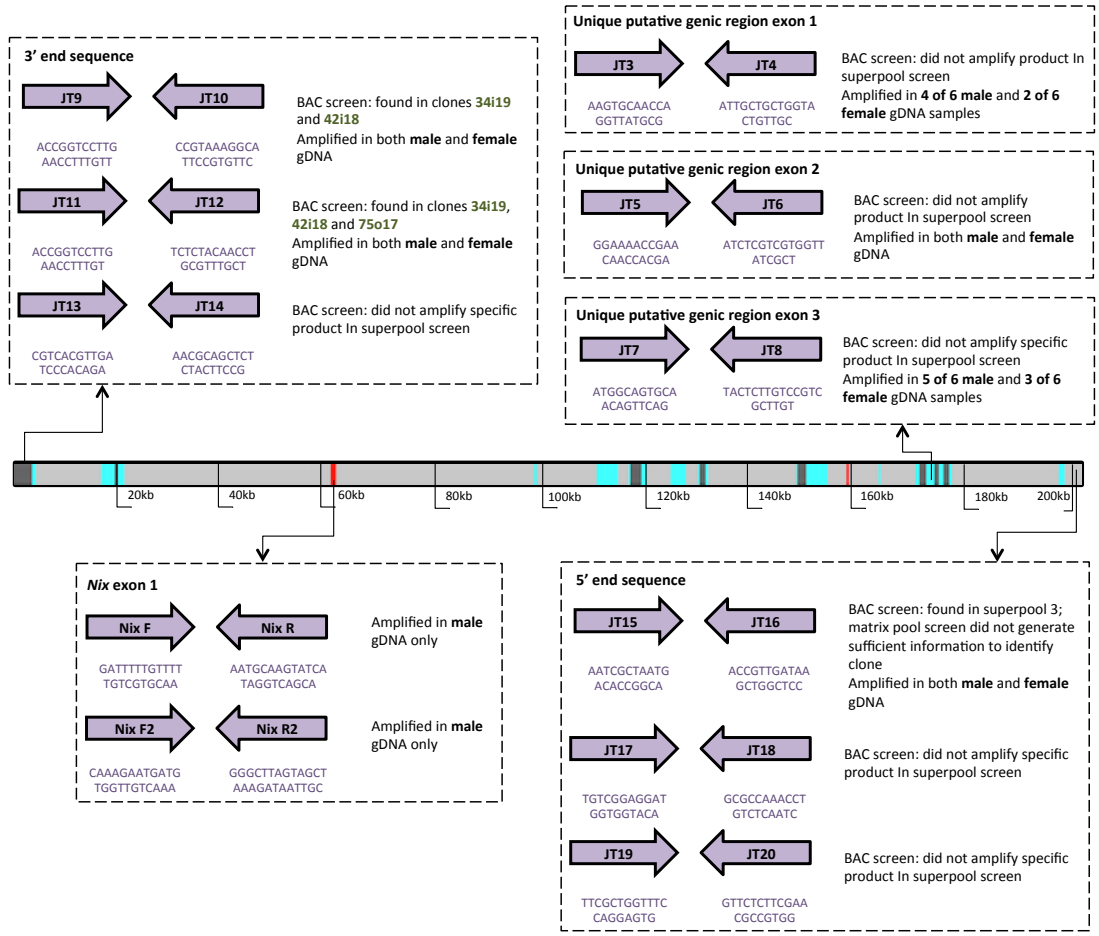


Figure 3.7 Results of BAC library and wild type gDNA PCR screening using primers targeting sequences in the 207kb Nix *region. Colours on the central track represent genomic features (red:* Nix *exons detected with BLASTN; cyan and dark grey: putative coding sequences and introns, respectively, predicted with Augustus (Keller* et al., *2011)).*

The BAC library was also screened for other sequences posited to be at or near the M locus. Of these, *myo-sex* and three candidate male-biased AaegL3 contigs gave positive hits for specific BAC clones and were detected predominantly in male rather than female wild type gDNA (Table 3.2).

Table 3.2 Results of BAC library and wild type gDNA PCR screening using primers targeting candidate male-biased sequences.

F primer	R primer	Sequence description	Present in BAC library clone(s)	Present in male gDNA (/6)	Present in female gDNA (/6)
JT9	JT10	*Nix* region 3' end sequence	34I19	6	6
JT11	JT12	*Nix* region 3' end sequence	34I19, 42I18, 75O11	6	6
JT39	JT40	*myo-sex*	16I12, 82B17	6	2
JT41	JT42	*myo-sex*	16I12, 82B17	6	2
JT43	JT44	*myo-sex*	–		
JT45	JT46	BAC NDL62N22	–		
JT47	JT48	BAC NDL62N23	–		
JT49	JT50	AY988440.1 18S rDNA	–		
JT51	JT52	AY988440.1 18S rDNA	–		
JT53	JT54	AY988440.1 18S rDNA	–		
M37	M38	AAGE02035037.1(1-6260)	–		
M41	M42	AAGE02035965.1(1-4650)	–		
M7	M8	AAGE02035016.1(1-6296)	103L20	6	0
M11	M12	AAGE02035557.1(1-5425)	72G8	6	0
SS1818	SS1819	AAGE02036067.1(1-4250)	6L17, 52E10, 77N2, 85C16, 101J3, 145C5	5	0
M1	M2	AAGE02035994.1(1-4545)	–		
M49	M50	AAGE02034767.1(1-6813)	–		

The 12 BACs were selected for PacBio sequencing: 6L17, 16I12, 34I19, 52E10, 72G8, 75O11, 77N2, 82B17, 85C16, 101J3, 103L20, 145C5. However, insufficient BAC DNA was acquired from the bacterial cultures, even after using larger culture volumes and an adapted protocol for very low-copy number plasmid extraction (Qiagen Plasmid Purification Handbook, April 2012).

3.5 Discussion

The genomic data from our assembled M locus region show that *Nix* is approximately 100 kb in length – exceptionally long even for an insect, and one of the longest in the mosquito genome. This is particularly unusual because *Nix* is expressed in early embryonic development, before the onset of the syncytial blastoderm stage 3-4 hours after oviposition (Hall *et al.*, 2015), during which time most active genes have very short introns, or lack them entirely. There is evidence of selection against intron presence in genes expressed in the early *Ae. aegypti* zygote (Biedler *et al.*, 2012). In *Drosophila*, the majority of early-expressed genes have small introns and encode small proteins, suggesting that selection has favoured high transcript turnover during early embryonic development due to the requirement for short cell cycles and rapid division (Artieri and Fraser, 2014). It might therefore be expected that selection would limit the *Nix* intron's expansion to preserve efficient transcription in the zygote.

One possible explanation is the expansion of repetitive DNA. The RepeatMasker results reveal that the *Nix* region contains a high number of repetitive sequences, especially retrotransposons (Figure 3.4; Table 3.1). The M locus has accumulated repeats in between protein-coding DNA in a manner characteristic of a sex chromosome, which are prone to degeneration by Muller's ratchet due to the lack of recombination (Muller, 1964; Charlesworth, 1991; Kaiser and Bachtrog, 2010). For instance, repetitive sequences comprise almost the entire *Anopheles gambiae* Y chromosome, and these repetitive sequences show rapid evolutionary divergence (Hall *et al.*, 2016), while the neo-Y chromosome in *Drosophila miranda* is considerably larger than the neo-X due to the widespread insertion of transposable elements in intergenic and intronic sequences (Mahajan *et al.*, 2018). Similarly, certain Y chromosome genes of the plant *Silene latifolia* have much larger introns than their X chromosome copies due to the insertion of retrotransposons (Marais *et al.*, 2008). A more extreme version of this phenomenon is seen in *Drosophila*

melanogaster, where some Y chromosome genes, such as those involved in spermatogenesis, have gigantic repetitive introns, sometimes in the megabase range, that consequently make them many times larger than typical autosomal genes (Carvalho *et al.*, 2001; Bachtrog, 2013).

It is therefore possible that the lack of recombination may pose constraints on the structure of the M locus, and in the absence of strong selection the *Nix* gene has degenerated outside the coding regions. This is also supported by the lack of specificity in the BAC library of many M locus sequences outside of the *Nix* exons. Non-coding sequences tend to be present in a high number of BAC clones and are occasionally amplified in the gDNA of female mosquitoes (Figure 3.7), and sections of the *Nix* region show high similarity to sequences across chromosome 1 (Figure 3.6B), indicating that copies of transposable elements present in the M locus occur throughout the genome.

Non-recombining sex loci such as the *Ae. aegypti* M locus may represent an evolutionary precursor to differentiated sex chromosomes, which are thought to emerge when sexually antagonistic alleles accumulate on either chromosome and favour reduced recombination between the two homologues, eventually leading to degeneration and loss of genes on the proto-Y (Charlesworth *et al.*, 2005). Recent data appears to show that recombination is reduced along autosome 1 even outside of the M locus (Fontaine *et al.*, 2017), while the fully differentiated *Anopheles* X and Y chromosomes still display some degree of recombination with each other (Hall *et al.*, 2016). Thus, *Ae. aegypti* may be "further along" this evolutionary trajectory than previously assumed. The presence of additional sequence in our BAC assembly, which was obtained from the AWT/My1 mosquito strain, compared to the corresponding region in the AaegL5 genome assembly obtained from the Liverpool strain – represented by a gap in the alignment of the two sequences (Figure 3.4; Figure 3.6) – suggests that the M locus may vary between strains outside of the *Nix* exons. Future work could investigate the population-level variation in the size and content of the M locus.

The *Ae. aegypti* M locus provides an intriguing example of the complexity of evolutionary forces acting on sex chromosomes, and further study of the locus will contribute to understanding the evolution of sex determination in insects and address general questions about the factors impacting gene and genome length. Importantly, these may also yield insights that can be applied to increase the efficiency of genetic strategies for vector control.

Chapter 4 Genomic analysis of the *Aedes aegypti* M locus

Part of this chapter has been published as

Matthews, B.J., Dudchenko, O., Kingan, S., Koren, S., Antoshechkin, I., Crawford, J.E., Glassford, W.J., Herre, M., Redmond, S.N., Rose, N.H., Weedall, G.D., Wu, Y., Batra, S.S., Brito-Sierra, C.A., Buckingham, S.D., Campbell, C.L., Chan, S., Cox, E., Evans, B.R., Fansiri, T., Filipovic, I., Fontaine, A., Gloria-Soria, A., Hall, R., Joardar, V.S., Jones, A.K., Kay, R.G.G., Kodali, V., Lee, J., Lycett, G.J., Mitchell, S.N., Muehling, J., Murphy, M.R., Omer, A., Partridge, F.A., Peluso, P., Aiden, A.P., Ramasamy, V., Rasic, G., Roy, S., Saavedra-Rodriguez, K., Sharan, S., Sharma, A., Smith, M., Turner, J., Weakley, A.M., Zhao, Z., Akbari, O.S., Black, W.C., Cao, H., Darby, A.C., Hill, C., Johnston, J.S., Murphy, T.D., Raikhel, A.S., Sattelle, D.B., Sharakhov, I. V, White, B.J., Zhao, L., Aiden, E.L., Mann, R.S., Lambrechts, L., Powell, J.R., Sharakhova, M. V, Tu, Z., Robertson, H.M., McBride, C.S., Hastie, A.R., Korlach, J., Neafsey, D.E., Phillippy, A.M., and Vosshall, L.B. (2018). Improved reference genome assembly of *Aedes aegypti* informs arbovirus vector control. *Nature.* **563**:501-507.

(See Appendix 1)

An earlier version of the manuscript was deposited on the preprint server *bioRxiv* and can be found at

https://doi.org/10.1101/240747

4.1 Abstract

In the arbovirus vector mosquito *Aedes aegypti,* male development is initiated by the gene *Nix*, which is limited to males because of its location within a non-recombining autosomal locus known as the M locus. Although the cDNA of *Nix* has been characterised, and its intron and the genome region in its immediate vicinity sequenced, little is known about the M locus region, except that it is located close to the centromere of one copy of chromosome 1 in males. The study of the M locus region is impeded by the small contig N50 and bias towards female sequences of the available mosquito reference genome. Analysis of the M locus is important both for understanding insect sex chromosome evolution and for the potential development of sex-targeted genetic control technologies. This chapter introduces the approach of the *Aedes* Genome Working Group (AGWG) to produce a more complete, high quality male genome assembly, AaegL5, and describes research conducted outside of the AGWG to use the improved assembly as a resource for investigating the M locus. AaegL5 contains nearly the full M locus sequence of approximately 1.5 Mb, and male-specific BACs confirmed its updated chromosome position at cytogenetic band 1p11. Further work showed that the ratio of alignments of male to female DNA reads was greatly increased over the M locus, which is present as a single haplotype, and the locus has a high density of retrotransposons (but no corresponding enrichment of piRNAs), typical of a young Y chromosome. The male-specific coverage extends further over chromosome 1, encompassing a region of approximately 80 – 100 Mb, indicating that recombination is reduced outside of the M locus, but any strict boundaries of suppressed recombination could not be precisely determined. Other than *Nix*, previously identified to be an M locus gene, no other additional male-limited genes could be found in the region. Interestingly, an orthologue of *Nix* is present in the closest relative of *Ae. aegypti* with a reference genome, *Ae. albopictus*. This orthologue shows high sequence similarity to the *Ae. aegypti* gene, suggesting that this gene may be the ancestral sex determining locus.

Future research could use the AaegL5 assembly as a reference to study these features in more detail.

4.2 Introduction

4.2.1 The importance of genomics in the study of mosquito biology

The increasing availability of sequencing technologies over the past several decades has resulted in reference genome assemblies becoming a vital part of biological research. While the limitations of short read technologies have posed a challenge to some aspects such as assembling heterochromatin, efforts to produce high quality genomes have enabled more precise and accurate analysis of a range of biological features of complex organisms (International Human Genome Sequencing Consortium, 2004). Over the past few years there has been a proliferation of sequencing platforms and computational strategies aimed at overcoming the challenges of assembling large genomes (Bradnam *et al.*, 2013). Advances in computer power and the plummeting costs of sequencing have led to efforts to generate genome assemblies for a wide range of organisms, the most ambitious of which involve sequencing all species in the UK (Stokstad, 2018), and even all eukaryotic organisms on Earth (Lewin *et al.*, 2018).

However, there are still obstacles to constructing optimal reference genomes. Research into genome architecture, large-scale structural variation, and the organisation of genes and chromatin requires chromosomal-length assemblies, which new methods such as linked reads and Hi-C are beginning to facilitate (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014). Furthermore, sequencing diploid genomes often results in collapsing alternative haplotypes into a single mosaic sequence that is not representative of the actual genome, which has led to the development of tools to phase alleles across assembled scaffolds (Korlach *et al.*, 2017; Koren *et al.*, 2018). Such information is crucial to the study of sex chromosome evolution, because sex determination factors are commonly present on one pair of undifferentiated chromosomes, such as in *Ae. aegypti.*

Due to its role as a vector of arboviruses, *Ae. aegypti* has been the focus of genomic studies aimed at improving disease control strategies, such as the study of genome-

wide variation in insecticide resistance genes across populations (Severson and Behura, 2012). In particular, analysing the sequence of sex chromosomes can uncover genes involved in sex determination, and has potential application to vector control (Adelman and Tu, 2016). However, the previously available reference genome, AaegL3 (Table 4.1), which is fragmented and biased towards female constituent sequences, does not contain the full M locus sequence (Hall *et al.*, 2014). Despite this impediment, knowledge of the M locus has increased in recent years, with studies identifying the principal M factor and deciphering its immediate genomic context (Hall *et al.*, 2015; Turner *et al.*, 2018), as well as exploring variation in the vicinity of the M locus (Campbell *et al.*, 2017; Fontaine *et al.*, 2017). However, detailed insights into the nature of this region, its evolutionary history, and the presence of additional M genes, are still hampered without a high quality reference assembly, especially one obtained from a male.

4.2.2 An improved, highly contiguous *Ae. aegypti* genome assembly

The limitations of the existing mosquito assembly inspired the creation of the *Aedes* Genome Working Group (AGWG), with the aim of producing a high quality genome map to facilitate detailed investigation into questions about the mosquito's biology (Harmon, 2016). The sequencing project used 80 sibling males of the Liverpool strain LVP_AGWG, which is closely related to the strain used for the previous AaegL3 assembly (Nene *et al.*, 2007), and sequenced the DNA on 177 PacBio RSII SMRT cells using P6-C4 chemistry. The PacBio data was assembled into a partially-phased diploid genome consisting of a total of 7,790 primary contigs as well as haplotigs (representing alternative alleles in the pool of individual males) with FALCON-UNZIP 0.7.0 (Chin *et al.*, 2016), followed by sequence polishing by mapping with BLASR 3.1.0 (Chaisson and Tesler, 2012) and consensus calling with QUIVER and ARROW (Chin *et al.*, 2013) (Table 4.1). After 359 contigs shorter than 20 kb were excluded, they were ordered, oriented and merged using Hi-C. The Hi-C data were used to analyse the frequency of contacts between loci in crosslinked

DNA, allowing scaffolding based on overlapping sequences between contigs and alternative haplotigs across large distances, and had previously been used to scaffold an improved, more contiguous *Ae. aegypti* genome assembly, AaegL4, than was available previously (Dudchenko *et al.*, 2017; Table 4.1).

After a final round of polishing and gap-filling with ARROW and PBJELLY 15.8.24 (English *et al.*, 2012), the resulting assembly (GenBank accession GCA_002204515.2) consisted of three chromosome-length scaffolds containing 94% of the sequenced bases, with the remaining 6% present in ~2,300 smaller scaffolds (Table 4.1). The assembly has an estimated base accuracy of 99.9665%, and still contains 229 gaps, of which 173 are present in the chromosome-length scaffolds (Matthews *et al.*, 2018).

Table 4.1 Assembly statistics for different Ae. aegypti *reference genome assemblies. Results are included for the updated AaegL5 male assembly before and after sequence polishing and Hi-C scaffolding. Table from Matthews* et al. *(2018).*

	AaegL3	**AaegL4**	**AaegL5 (FALCON-Unzip)**	**AaegL5 (NCBI)**
Total length (bp)	1,310,092,987	1,254,548,160	1,695,064,654	1,278,709,169
Contig number	36,205	37,224	3,967	2,539
Contig N50 (bp)	82,618	84,074	1,304,397	11,758,062
Scaffold number	4,757	6,206	–	2,310
Scaffold N50 (bp)	1,547,048	404,248,146	–	409,777,670
GC content (%)	38.27	38.28	38.16	38.18
Alternative haplotig number	–	–	3,823	4,224
Alternative haplotigs (bp)	–	–	351,566,101	591,941,260

Illumina RNA-Seq and PacBio IsoSeq data was applied to the NCBI RefSeq annotation pipeline (Thibaud-Nissen *et al.*, 2013) to generate a more complete gene set, AaegL5.0 (RefSeq accession GCF_002204515.2). Analysis of annotation completeness using single copy universal orthologues with BUSCO (Simão *et al.*, 2015) found there were fewer fragmented and missing genes than previous

assemblies (Table 4.2). With the use of Bionano optical mapping, 10x linked read sequencing, and fluorescent *in situ* hybridisation (FISH), potential misassemblies could be corroborated, and structural variants within the LVP_AGWG strain could be detected (Matthews *et al.*, 2018).

Table 4.2 BUSCO results for different Ae. aegypti *reference genome assemblies, representing the completeness of a set of known universal single-copy orthologues within each assembly. Duplicated genes indicate potential alternative haplotypes in each assembly. Data from Matthews* et al. *(2018).*

	AaegL3	**AaegL4**	**AaegL5 (FALCON-Unzip)**	**AaegL5 (NCBI)**
Complete	96.4%	95.4%	97.7%	96.7%
Single-copy	91.1%	93.1%	46.3%	93.0%
Duplicated	5.3%	2.3%	51.4%	3.7%
Fragmented	2.0%	2.4%	1.1%	1.8%
Missing	1.6%	2.2%	1.2%	1.5%

4.2.3 Background and chapter aims

Matthews *et al.* (2018) show that the improved reference genome AaegL5 enables a detailed study of a variety of features of *Ae. aegypti*, including the discovery of additional odorant receptors, measurement of genome-wide genetic variation, and identification of quantitative trait loci involved in pyrethroid resistance and vector competence for dengue. The alignment of AaegL5 to the AaegL4 assembly with LAST (Kiełbasa *et al.*, 2011) identified a divergent region around the locations of the M locus genes *Nix* and *myo-sex*, and a matching scaffold in the Bionano optical map assembly that spanned this region was identified with BLASTN (Altschul *et al.*, 1990), allowing the approximate boundaries of a putative ~1.5 Mb M locus to be determined (Matthews *et al.*, 2018). Notwithstanding a gap of approximately 163 kb between two contigs, this represents the most complete sequence of the *Ae. aegypti* M locus ascertained so far.

This chapter describes the work carried out as part of the AGWG leading to the placement of the M locus genes using the BACs identified in Chapter 3, as well as research undertaken outside the AGWG, using the male assembly as a reference, to

complement and expand the findings from the original study and discover more information on the genome architecture relating to the M locus. This includes: 1) analysing male and female short DNA read coverage across the entire genome to look for regions of sex-differentiation; 2) aligning 10x linked reads from a single male to identify translocations and inversions associated with the M locus; and 3) examining these data to pinpoint the boundaries of the M locus, via evidence of regions of reduced recombination and male-specific heterozygosity adjacent to *Nix*. As sex-linked, non-recombining regions often show an enrichment of repetitive DNA, the abundance and distribution of different types of transposable elements in the M locus was investigated, as well as that of small RNAs (smRNAs) that are often involved in silencing transposons (Malone and Hannon, 2009). In addition, putative additional genes in the M locus were identified by assembling a *de novo* transcriptome and eliminating non-male transcripts. Finally, differential male-female coverage and repeat content around the M locus orthologues in *Ae. albopictus* was investigated to determine whether male specific sequences displayed similarity between the two related species.

4.3 Materials and Methods

4.3.1 Preparation of M locus BACs for FISH

The four *Nix* BAC clones and 12 additional candidate male BAC clones described in Chapter 3 were plated onto selective chloramphenicol agar plates and incubated overnight at 37°C. Individual colonies were picked from each plate using sterile toothpicks and placed into cryogenic vials (Starlab, UK) approximately two thirds filled with LB agar. These stab cultures were sent to the Sharakhova Lab at Virginia Tech, VA, USA, to act as templates for probes to perform FISH according to published protocols (Sharakhova *et al.*, 2011; Timoshevskiy *et al.*, 2013).

4.3.2 Analysis of differential male and female coverage of sequencing data

4.3.2.1 Sequencing and preliminary work

The Centre for Genomic Research (CGR), Liverpool, prepared PCR-free Illumina libraries from the genomic DNA of 12 male and 12 female *Ae. aegypti* of the inbred Asian wild type (My1) Family 2 strain described in Chapter 2.3.1.1, which were sequenced on two lanes of the Illumina Hi-Seq platform as 2 x 150 bp paired end reads (Illumina, USA). Sequencing adapters were trimmed and low quality bases were removed. These DNA read data are henceforth referred to as the CGR dataset. Other datasets were downloaded from the Sequence Read Archive (SRA) (referred to as Virginia Tech, Rockefeller and Cambridge datasets) and are detailed in Appendix 2.1.

As discussed above in Chapter 2, Ritesh Krishna developed a pipeline for identifying sex-biased contigs in the AaegL3 reference assembly by examining the differential coverage of aligned male and female reads. This technique was modified and applied to the updated AaegL5 assembly.

4.3.2.2 Differential coverage analysis to detect sex-biased sequences

The reads from each of the 24 libraries were mapped to the assembly with BOWTIE 2.2.1 (Langmead and Salzberg, 2012) using default parameters with -I 200 and -X 500. The alignments were merged to create single BAM files for all males and all females, and these were sorted by coordinate and indexed with SAMTOOLS 0.1.8 (Li *et al.*, 2009a). The alignments were filtered with SAMTOOLS to only retain mappings with a quality of 10 or higher, in order to remove alignments to multiple locations such as transposable elements. Potential PCR duplicates were removed using the MarkDuplicates tool of Picard (Broad Institute, broadinstitute.github.io/picard). Coverage statistics for the merged male and female alignments were calculated with BEDTOOLS 2.16.2 (Quinlan and Hall, 2010) using the coverageBED subcommand. The BED file produced by the BEDTOOLS coverageBED subcommand contains two coverage metrics: the number of reads that mapped with at least one base pair to the reference, referred to here as the *depth* of coverage; and the fraction of bases in the reference that had reads mapped, referred to here as the *breadth* of coverage. Thus the coverage of male and female reads across a given interval in the reference can be theorised in two dimensional space, with the value d representing the difference between the combined depth and breadth of male and female data, and therefore the sex-specificity of the interval (Figure 4.1).

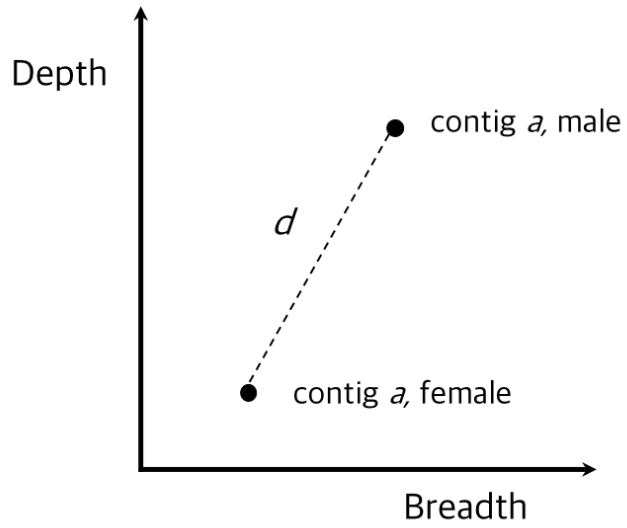


Figure 4.1 The relative depth and breadth of coverage of DNA reads from male and female datasets across a hypothetical interval in the reference genome (contig a*). The value* d *represents the discrepancy between the combined coverage metrics for each dataset, and is therefore a measure of the sex-specificity of the interval.*

A MATLAB script (Appendix 3.4.5) was used to calculate *d* for each interval in the reference genome and return those with the greatest sex bias (i.e. those with a high ratio of alignments from male reads compared to female reads). The candidate intervals were examined by eye with the Integrative Genomics Viewer (Robinson *et al.*, 2011a) to validate the presence of more male alignments than female.

4.3.2.3 Chromosome Quotient analysis

A similar method of detecting sex-biased sequences was developed by Andrew Brantley Hall, named the Chromosome Quotient (CQ) method (Hall *et al.*, 2013). The method uses BOWTIE 0.12.7 (Langmead *et al.*, 2009) to separately map male and female Illumina data to reference sequences and calculate the ratio of female-to-male alignments. Sequences with a low ratio – those with alignments from male data but very few from female data – are potentially associated with male-specific regions such as Y chromosomes or the M locus.

The advantage of this method is that male-specific sequences with some degree of alignment from female data, such as Y chromosome genes with closely-related

paralogues on the X chromosome, are not discarded as candidates. Furthermore, as a default, the CQ method requires at least 20 alignments from male data for any sequence to be reported, thereby avoiding false positives due to misassemblies or bacterial contamination in reference sequences. Consequently, the CQ method is more reliable than methods based on eliminating sequences with female alignments (Carvalho and Clark, 2013), and has been demonstrated to successfully identify male-specific sequences in the mosquito species *Ae. aegypti* and *An. gambiae* (Hall *et al.*, 2014; Hall *et al.*, 2015; Hall *et al.*, 2016).

To compare and corroborate the results of the analysis in 4.3.2.2 above, the CQ script (github.com/brantleyhall/Chromosome-Quotient) was run on the above male and female Illumina data, using default parameters, with the reads from individual runs concatenated to give single male and female FASTQ files.

4.3.3 Analysis of structural variance using 10x linked reads

4.3.3.1 Phenol-chloroform extraction of high molecular weight DNA

Ae. aegypti from the AWT Family 2 strain were reared according to the protocol described in Chapter 2.3.1.2. A single male pupa was ground in liquid nitrogen in a 2 ml Eppendorf (Eppendorf, Germany) and 1 ml TLB (50% Tris-Cl, 25% SDS, 20% EDTA) was added along with 5 µl proteinase K (Qiagen, Germany), and the mixture incubated at 50°C, 300 rpm for 1 h. 5 µl RNAse A (Qiagen, Germany) was added and incubated at 37°C, 300 rpm for 1 h.

The lysate was poured into a 15 ml Falcon tube containing Phase Lock Gel light aqueous (Quantabio, USA) and 1 ml phenol (Sigma-Aldrich, USA) was added. The mixture was mixed at 4°C, 40 rpm for 10 min and then centrifuged at 4°C, 4000 rpm for 12 min. The aqueous layer remaining above the gel was poured into a new phase lock gel tube and an equal volume of chloroform-isoamyl alcohol (Sigma-Aldrich, USA) added, mixed at 4°C, 40 rpm for 10 min and centrifuged at 4°C, 4000 rpm for 12 min.

The aqueous layer was poured into a 5 ml Eppendorf and 2 ml chilled (-20°C) ethanol and 300 µl sodium acetate (Sigma-Aldrich, USA) was added. The mixture was incubated at -80°C for 10 min and centrifuged at 4°C, 7000 rpm for 10 min. The supernatant was added to 1 ml 70% ethanol and centrifuged at 4°C, 10,000 x *g* for 1 min. The supernatant was removed with a pipette, and the pellet was air dried at 40°C for 10 min before being resuspended in 50 µl nuclease-free water (ThermoFisher, UK).

The DNA was quantified on a Qubit fluorometer using the dsDNA HS Assay Kit (ThermoFisher, UK), and run on a 0.5% agarose gel at 30V for 16 h to confirm large fragment size.

4.3.3.2 Library preparation and sequencing

The extracted DNA was submitted to the CGR, Liverpool, UK for library preparation and the sample sequenced by Edinburgh Genomics, Edinburgh, UK. Amplicons were barcode tagged on a single lane of the 10x Chromium instrument (10x Genomics) and sequenced on the Illumina HiSeq X10 as 2 x 150 bp paired end reads.

4.3.3.3 Data analysis

Edinburgh Genomics demultiplexed the samples and converted the barcode and read data to FASTQ files. The reads were mapped to the AaegL5 reference genome assembly with LONG RANGER 2.1.6 (10x Genomics) with the wgs subcommand, using default parameters and the option --sex=male. Only the three chromosome-length scaffolds of the AaegL5 genome were included in the reference because LONG RANGER allows only 500 reference sequences. The output was examined in LOUPE (10x Genomics) to inspect the structural variance and haplotype phasing across the genome.

The FASTQ files were also assembled with SUPERNOVA (10x Genomics), with default parameters and the option --style=pseudohap for the mkoutput

subcommand step. The assembly quality was assessed with QUAST 4.6.3 (Gurevich *et al.*, 2013).

4.3.4 Analysis of abundance and types of repeats

The AaegL5 reference genome assembly was analysed with REPEATMASKER 4.0.6 (Smit *et al.,* repeatmasker.org) using the *Ae. aegypti* RepBase library with the crossmatch algorithm set to 'sensitive', skipping the check for bacterial sequences, with a maximum sequence length of 20,000 and a cutoff score of 255. The tool ONE CODE TO FIND THEM ALL (Bailly-Bechet *et al.*, 2014) was run with default parameters on the REPEATMASKER output to consolidate information on the positions and abundances of types of repeats across the genome.

4.3.5 Analysis of small RNAs

smRNA data from male and female *Ae. aegypti* was downloaded from the SRA (accession numbers SRR1068553, SRR5961505, SRR5961506; Appendix 2.1) and mapped to the chromosome-length scaffolds of the AaegL5 reference genome assembly with BOWTIE 2.2.1 (Langmead and Salzberg, 2012) using the --fast preset, which reports only the best matched alignment for smRNAs mapping to multiple locations (Lewis *et al.*, 2018). Coverage statistics were calculated for the male and female data in 100kb bins across the genome using BEDTOOLS 2.16.2 coverageBED (Quinlan and Hall, 2010).

4.3.6 Identification of additional candidate M locus genes

An approach was used to find male-limited genes other than *Nix* and *myo-sex* in the M locus, based on a pipeline previously developed to identify Y chromosome-linked coding sequences in mammals and birds (Cortez *et al.*, 2014), and later applied to Dipteran species (including *Ae. aegypti*, for which it was unsuccessful) (Mahajan and Bachtrog, 2017). The pipeline progressively eliminates transcripts from a male-specific transcriptome that show similarity to female sequences, resulting in putative

male-specific genes such as those only present in male genome regions like the Y chromosome or M locus.

Male RNA-Seq data (SRA accession SRR924021; Appendix 2.1) were mapped to the "female" reference *Ae. aegypti* genome assembly AaegL4 (Dudchenko *et al.*, 2017) with TOPHAT 2.1.1 (Kim *et al.*, 2013) using default parameters. Reads that failed to map to the assembly were filtered out with SAMTOOLS (Li *et al.*, 2009a) and assembled into a *de novo* transcriptome with TRINITY 2.6.5 (Grabherr *et al.*, 2011) using default parameters. The assembled transcripts were mapped to the AaegL4 assembly with BLAT v.34 (Kent, 2002) with the minimum score set to 50 and minimum identity set to 98%, using the -fine option.

Transcripts that mapped to the assembly were discarded using BBTOOLS v.15 (Bushnell, sourceforge.net/projects/bbmap/). Female RNA-Seq reads (SRR1585315; Appendix 2.1) were mapped to the remaining transcripts with BOWTIE 2.2.1 (Langmead and Salzberg, 2012) using default parameters. Coverage statistics for each transcript were calculated using BEDTOOLS 2.16.2 coverageBED (Quinlan and Hall, 2010), and transcripts with more than 50% breadth of coverage with up to two mismatches were filtered out with BBTOOLS v.15.

The remaining transcripts were merged with TGICL 2.1 (Pertea *et al.*, 2003) using a minimum overlap of 30 bp to attempt to produce complete transcripts. Merged male and female CGR Illumina reads were mapped separately with BOWTIE 2.2.1 using default parameters and the coverage of male and female data for each transcript was calculated with SOAPCOVERAGE 2.7.7 (Li *et al.*, 2009b). Transcripts with a male breadth of coverage of less than 60% and female breadth of coverage of more than 10% were discarded.

Male and female RNA-Seq reads were mapped separately to the remaining transcripts with BOWTIE 2.2.1 using default parameters and expression values (reads per kilobase million; RPKM) were calculated using KALLISTO 0.43.1 (Bray *et al.*, 2016). Transcripts with less than a 2:1 male-to-female expression ratio were discarded.

To filter out transcripts that might stem from transposable elements, the *Ae. aegypti* repeat library was downloaded from VectorBase (Giraldo-Calderón *et al.*, 2015) and transcripts were mapped to the library with BLAT v.34 with default parameters. Transcripts that did not match any repeats were considered candidate male transcripts and were searched for across *Ae. aegypti* reference genome assemblies with BLASTN (Altschul *et al.*, 1990).

4.3.7 Comparison with the M locus in the mosquito *Aedes albopictus*

Ae. albopictus data was generated as part of a separate study investigating mosquitoes that were infected or cured with *Wolbachia*. The strains originated from Hawaii (Haw), Malaysia (KLP and Mal) and La Reunion (LR) and have been held at Oxitec Ltd since 2006 (Mal), 2010 (Haw and KLP) and 2012 (LR) (A. Darby, personal communication). DNA was extracted from pools of 10 cured or infected mosquitoes, PCR enriched libraries were prepared, and multiplex pooled samples were sequenced on the Illumina HiSeq4000 as 2 x 150 bp paired-end reads. Cured and infected samples were run on separate Illumina lanes, but this was disregarded and only the sex of the samples was considered for this chapter.

The data was applied to the differential coverage and CQ pipelines detailed above in sections 4.3.2.2 and 4.3.2.3 respectively, using the C6/36 cell line *Ae. albopictus* genome assembly as a reference (Miller *et al.*, 2018) (GenBank accession GCA_001876365.2). The abundance and distribution of transposable elements was analysed with RepeatMasker 4.0.6 (Smit *et al.,* repeatmasker.org) and one code to find them all (Bailly-Bechet *et al.*, 2014) similarly to the method detailed above in section 3.4, using the *Ae. albopictus* repeat library from VectorBase (Giraldo-Calderón *et al.*, 2015).

4.4 Results

4.4.1 The cytogenetic location of the M locus

Atashi Sharma in the Sharakhova Lab, Virgina Tech, performed FISH using probes constructed from the *Nix*- and *myo-sex*-containing BACs. The resulting images show that both probes hybridise to one homologous copy of chromosome 1 (Figure 4.2). Interestingly, the genes colocalised to band 1q11, which is different to previous studies that placed both genes at band 1q21 (Hall *et al.*, 2014; Hall *et al.*, 2015); however, the location is consistent with previous cytological and genetic linkage studies that deduced the location of the M locus to be in the pericentromeric region of chromosome 1 (Bhalla and Craig, 1970; Newton *et al.*, 1974; Newton *et al.*, 1978).

Figure 4.2 FISH on mitotic chromosomes of male Ae. aegypti *using probes containing* Nix *and* myo-sex*, indicating the location of the M locus within the cytogenetic band 1p11. Scale bar is 2 µm. Figure by A. Sharma.*

4.4.2 Differential male-female coverage analysis

4.4.2.1 Male-biased regions across chromosome 1

As expected, the ratio of female to male coverage of genomic DNA reads shows a considerable drop on chromosome 1 of the AaegL5 reference genome assembly around the M locus (Figure 4.3). This is particularly pronounced in the region between 151.68 – 152.95 Mb, which contains *Nix* and *myo-sex*, both of which Matthews *et al.* (2018) hypothesise to be in the M locus (Figure 4.4). However, the depressed female to male coverage on chromosome 1 extends beyond the M locus proper, comprising a region of approximately 80 Mb, a pattern that is not found on other chromosomes (Figure 4.3 and Figure 4.4). A similar pattern is found for other Illumina datasets aligned to AaegL5, with some variation in the extent of the male-biased region of chromosome 1 (Supplementary Figure 1; the Cambridge dataset contained a substantially greater number of female DNA samples than male, explaining the generally higher female coverage outside the M locus). Further evidence for this is found by comparing the male and female coverage of windows across the three chromosomes, with many windows from chromosome 1 exhibiting sex-biased coverage, while the male and female coverage is more tightly correlated on the other two chromosomes (Figure 4.5).

The extended region of male-biased coverage on chromosome 1 also contains the most male-specific sequences from the previous genome assembly, AaegL3. The 35 AaegL3 contigs identified as having much higher coverage of male than female reads (described in Chapter 2.2.5) – including the three selected for transformation with CRISPR/Cas9 in Chapter 2 and the additional four identified in the BAC library in Chapter 3 – were aligned to the AaegL5 assembly with BLASTN. For all contigs, the top hit was found to be approximately between 170 Mb – 200 Mb on chromosome 1 (Figure 4.6; Supplementary Table 1). This suggests that while some sex-differentiated sequences are encompassed in the original genome assembly, none of the ~1.5 Mb M locus included in the new assembly was present in the original.

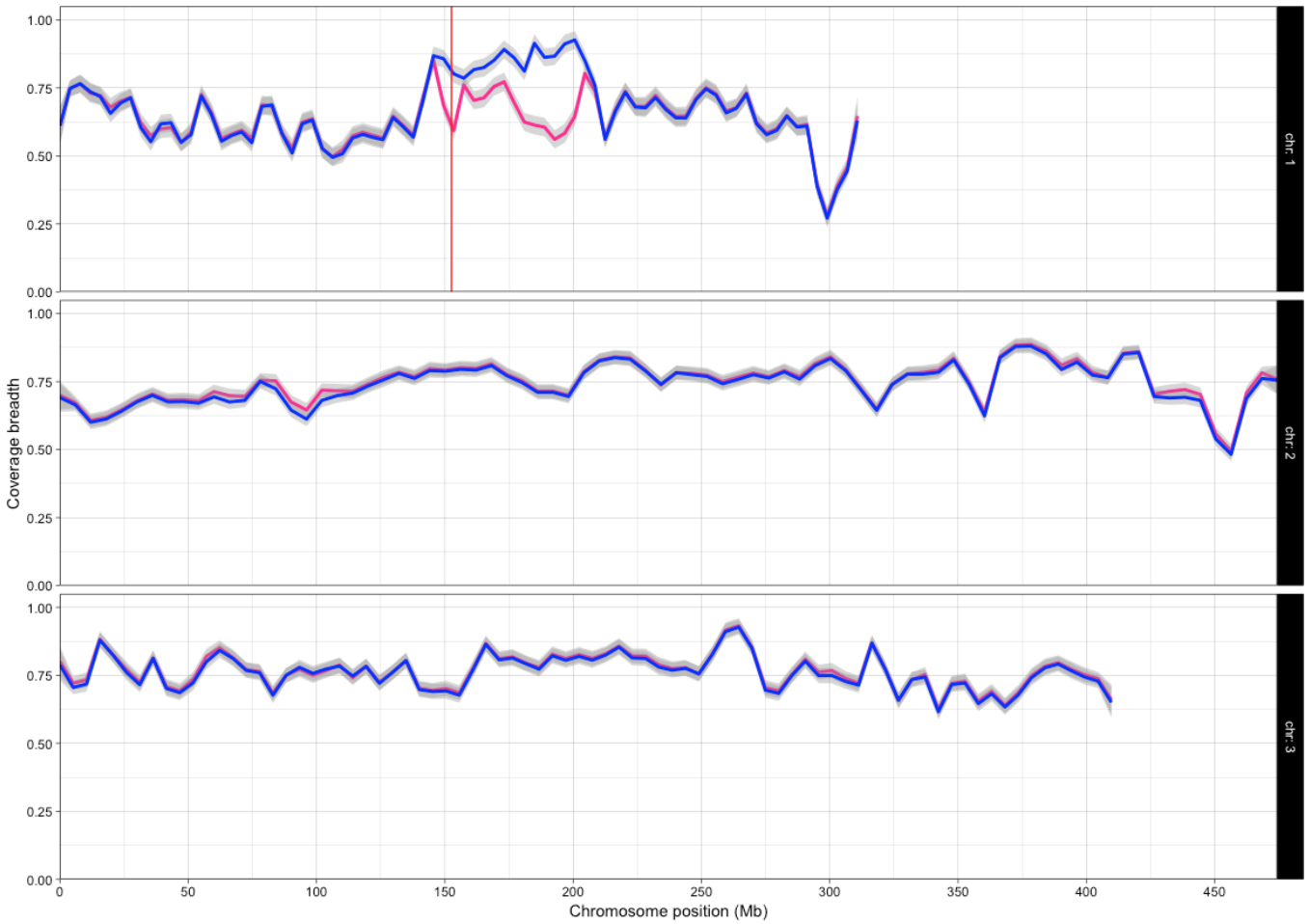


Figure 4.3 Breadth of coverage of female (red) and male (blue) genomic reads across 30 kb bins throughout the AaegL5 Ae. aegypti *reference genome assembly. The lines are smoothed conditional means; grey shading around the lines represents 95% confidence intervals. The red vertical line indicates the position of the M locus. Chromosomes are different sizes, resulting in data shown only for part of the window for chr 1 and 3.*

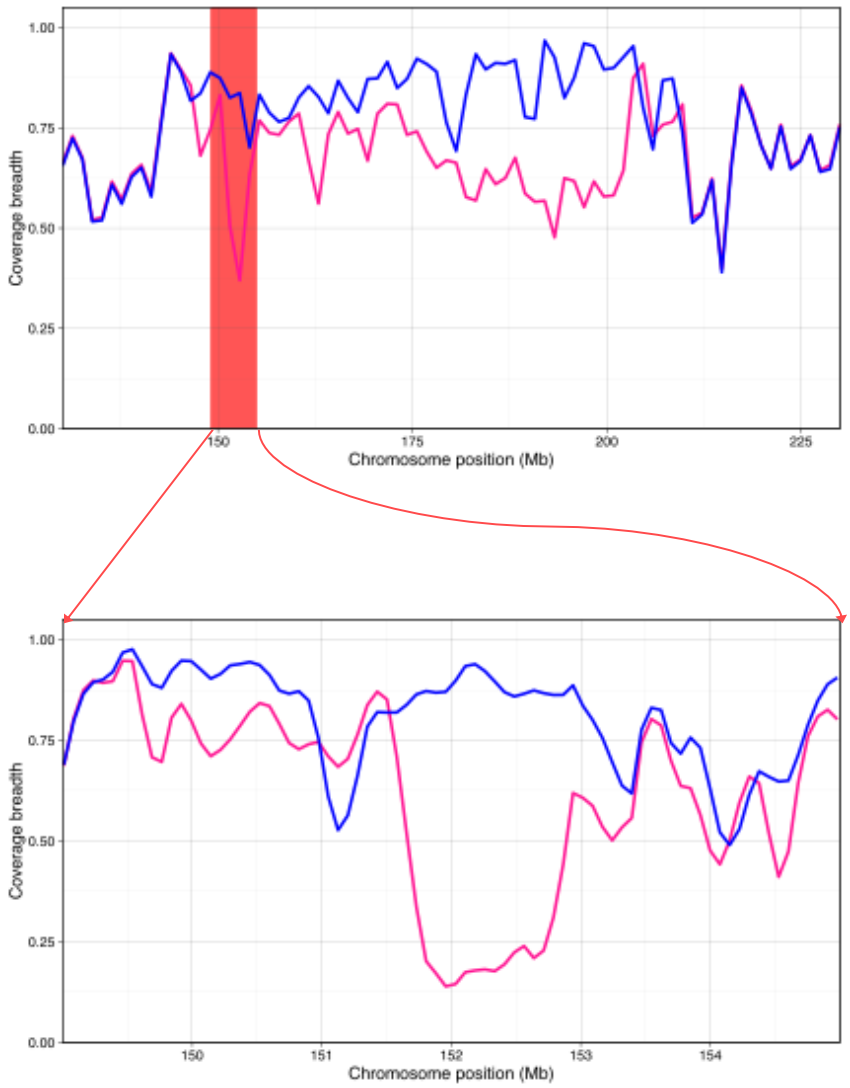


Figure 4.4 Breadth of coverage of female (red) and male (blue) genomic reads across 30 kb bins over sections of chromosome 1 of the AaegL5 Ae. aegypti *reference genome assembly. The top panel shows the region of male-biased coverage and the bottom panel shows the zoomed-in M locus. The lines are smoothed conditional means so the values in the two panels are different at their respective resolutions.*

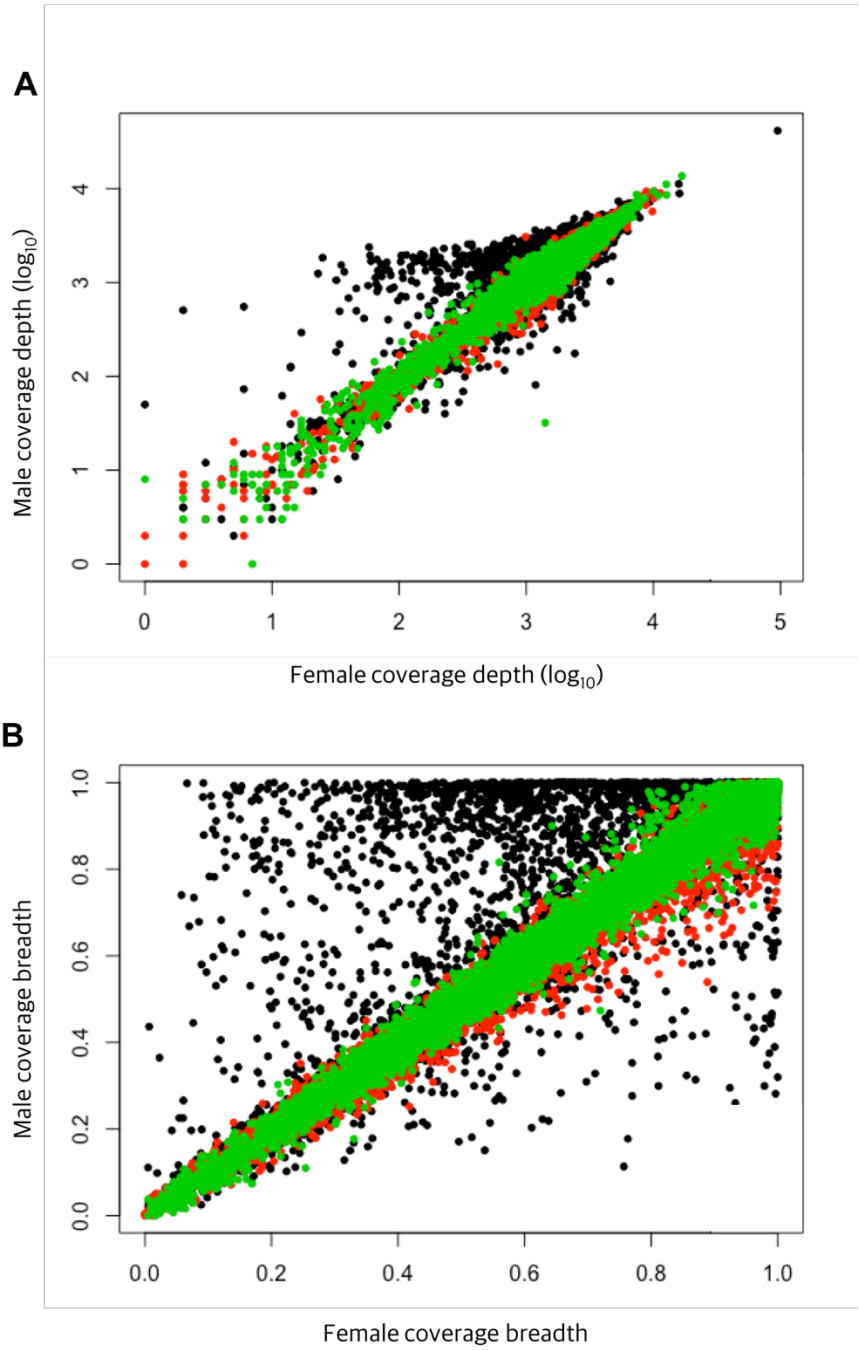


Figure 4.5 ***A*** *log_{10} depth of coverage and* ***B*** *breadth of coverage of female and male genomic reads across the AaegL5* Ae. aegypti *reference genome assembly. Each point represents a 30 kb bin. Colours represent the chromosome on which each bin is located (black: chr 1; red: chr 2; green: chr 3).*

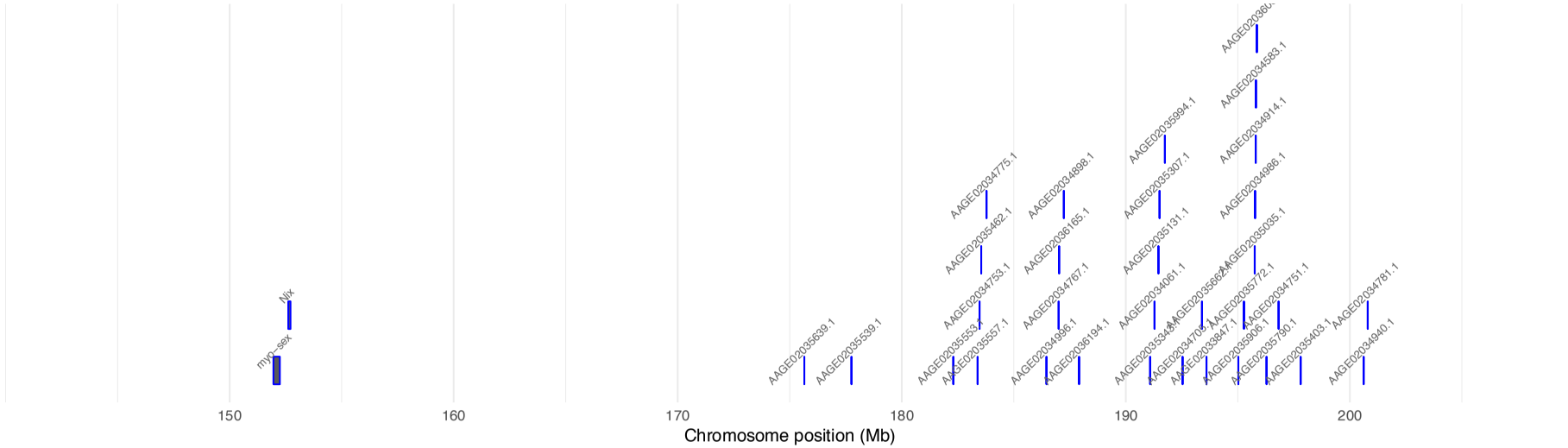


Figure 4.6 BLAST alignments of the 35 most male-biased contigs in the AaegL3 reference genome assembly, identified using a differential mapping coverage pipeline, against chromosome 1 of the AaegL5 male assembly, relative to the locations of the M locus genes Nix *and* myo-sex.

4.4.2.2 Comparison with CQ method

A comparable but much less pronounced pattern to the differential coverage analysis was observed in the CQ analysis (Figure 4.7). The median CQ score across the genome is 1.62 while it is 1.11 at a finer scale across the M locus, demonstrating that the ratio of female to male alignments is lower at the M locus, but still slightly biased towards female coverage. This is higher than expected, as CQ analysis performed with different male and female Illumina data resulted in a distinct region of very low CQ scores across the M locus (Matthews *et al.*, 2018). The reference sequence is typically repeat masked for the CQ analysis, but was not for this analysis; however, repeating the pipeline on the masked AaegL5 genome resulted in a median CQ score of 2.16 across the genome, indicating that alignment of female reads to repetitive sequence was not responsible for the female bias. An unequal number of reads in the male and female datasets can result in biased CQ scores if they are not normalised, however the number of reads is only 1.10 times greater in female dataset (94,222,660 read pairs compared to 85,622,222 in the male dataset).

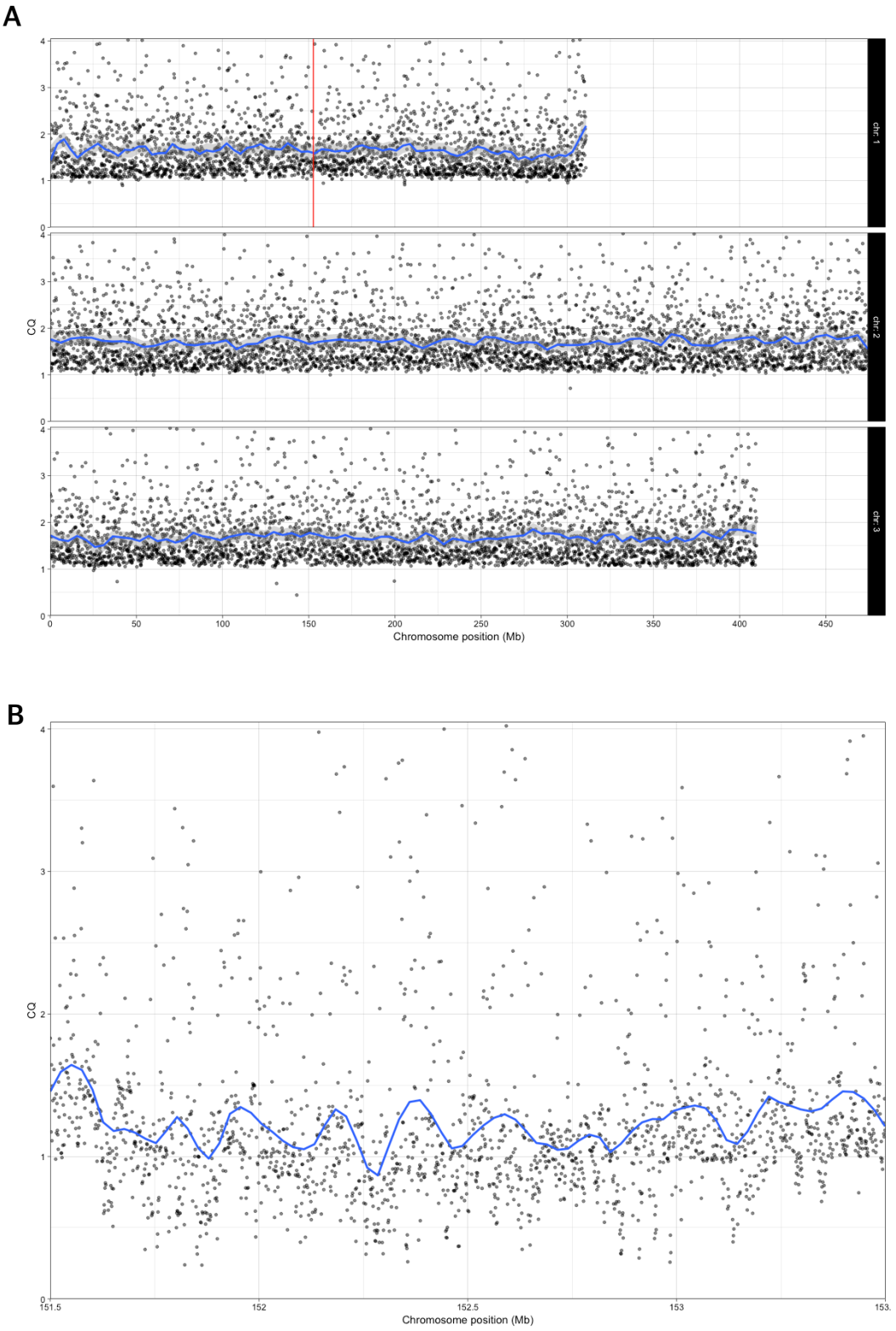


Figure 4.7 CQ values for ***A*** *100 kb bins across the AaegL5* Ae. aegypti *reference genome assembly; and* ***B*** *1 kb bins across the M locus on chromosome 1. The blue line is the smoothed conditional mean. The red vertical line in* ***A*** *indicates the position of the M locus.*

4.4.3 10x linked reads

The alignment of the male 10x linked reads to the AaegL5 genome shows that the great majority of variation is heterozygous over the M locus, which is located in one contiguous phased block, strengthening the evidence that it is only present on one copy of chromosome 1 (Figure 4.8). Structural variant analysis found two deletions of several kilobases in the linked read data upstream and downstream of the M locus relative to the reference genome (Figure 4.9), however no inversions or translocations were detected in the vicinity of the M locus.

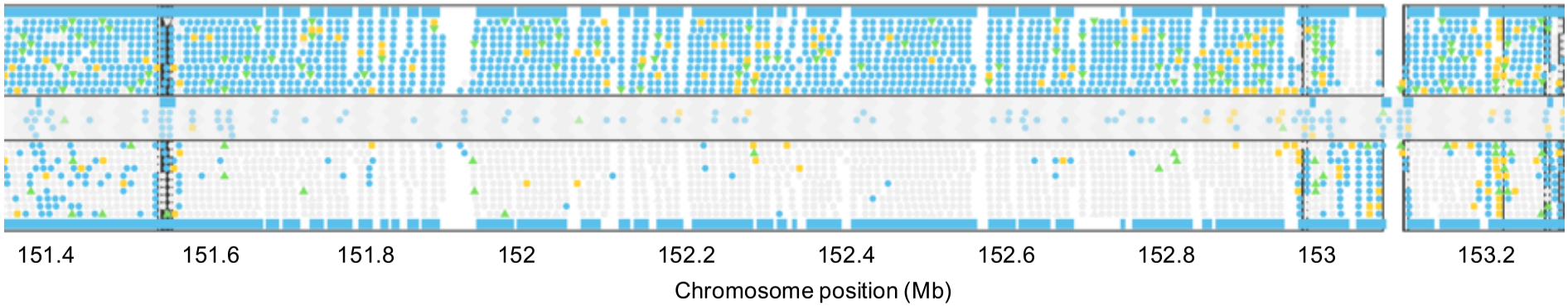


Figure 4.8 Haplotypes across the M locus on chromosome 1 of the AaegL5 reference genome assembly, determined by alignment of male linked reads from the AWT Family 2 strain with the 10x Long Ranger software. Icons display the location of phased (top and bottom tracks) and unphased (centre track) SNPs. Black boxes indicate phased blocks, with the region between ~151.5 – 153.1 Mb encompassed in a single block.. Icons represent different classes of SNPs (blue circle: substitution; green triangle: insertion; yellow square: deletion).

The linked reads assembled into a genome comprised of 81,620 contigs (Table 4.3). *Nix* and *myo-sex* were determined to be present in the assembly using BLASTN, located on two and four separate contigs, respectively (Supplementary Figure 2). However, the assembly was not contiguous enough to span the ~163 kb gap in the AaegL5 M locus sequence. An attempt was made to reassemble the full M locus by extracting reads from the LONG RANGER alignment that mapped to the AaegL5 M locus with one or both pairs, but too few of these reads were tagged with valid barcodes for SUPERNOVA to successfully assemble. Instead, the reads were assembled with SPADES 3.11.1 (Bankevich *et al.*, 2012) using the --only-assembler option, but this also resulted in a fragmented assembly (~1.66 Mb in 655 contigs, with an N50 of 5,420 bp) that could not be used to determine the content of the M locus gap.

Table 4.3 Assembly statistics for the male AWT Family 2 Ae. aegypti *genome generated from linked reads with the 10x Supernova software.*

Contig number	Largest contig (bp)	Total length (bp)	N50 (bp)	GC (%)
81,620	26,830,579	1,822,277,331	559,972	38.11

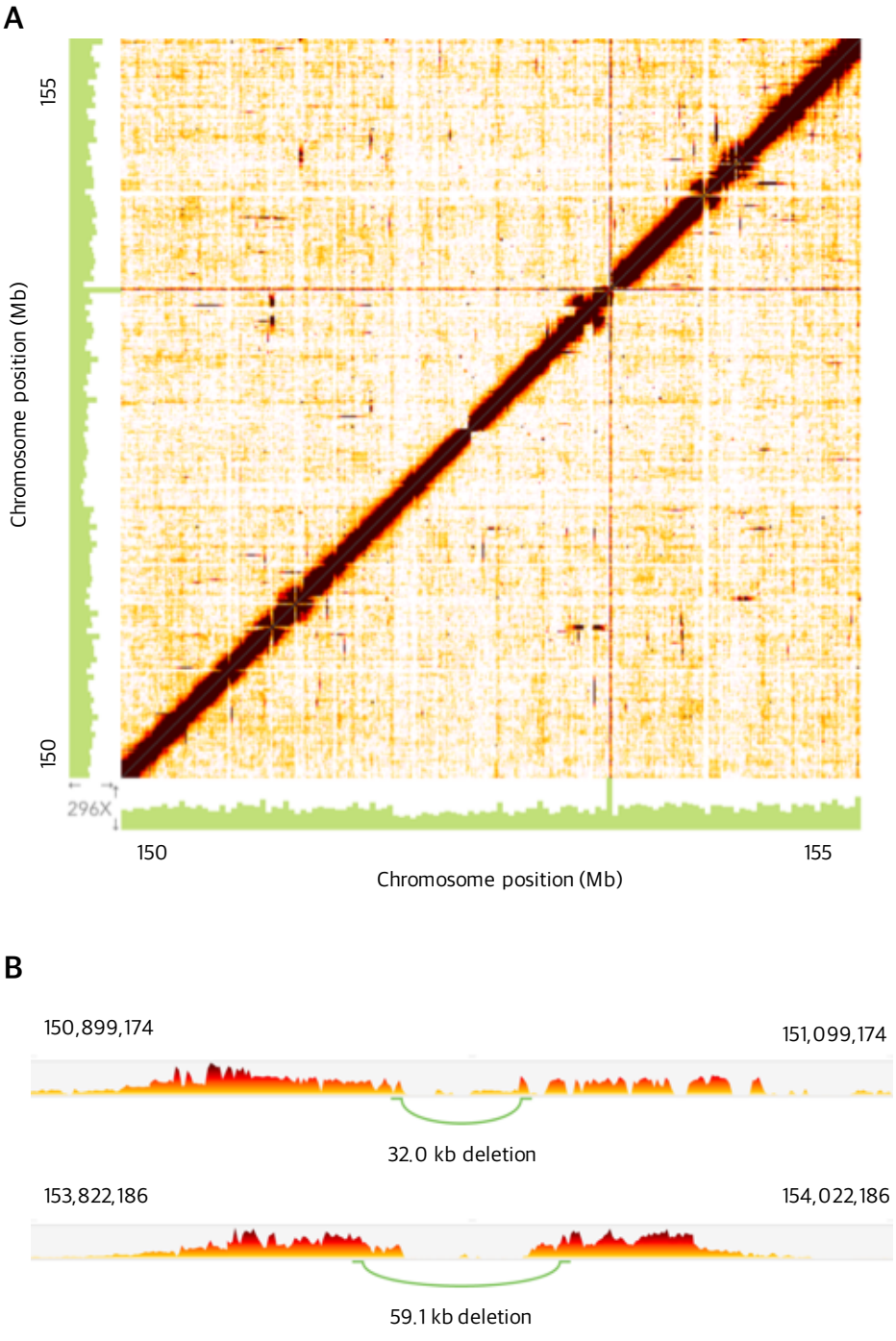


Figure 4.9 Visualisation of structural variation across a 5 Mb region of chromosome 1 in the AaegL5 reference genome assembly, identified with the 10x LongRanger software. ***A*** *The density of overlapping barcodes from reads aligned across the region; colour intensity represents the number of overlapping barcodes from both loci. Green bars show per base coverage across the region.* ***B*** *Linear view of two large deletions (green arcs) in the linked reads relative to the reference.*

4.4.4 Abundance and types of repeats

Transposable elements are abundant across all three chromosomes of the AaegL5 reference genome assembly, with a mean coverage of approximately 45%, consistent with the previous AaegL3 reference assembly (Nene *et al.*, 2007). The abundance of TEs tends to increase around the centromere in all three chromosomes, and this is particularly pronounced for chromosome 1 (Figure 4.10). The increased abundance appears to be mainly due to the presence of LTR retrotransposons, which are particularly enriched in the region in comparison to other classes of TEs such as DNA transposons and LINEs (Figure 4.11). On chromosome 1, the increased repetitive content is primarily explained by an enrichment of Gypsy and Copia LTR elements (Figure 4.12).

4.4.5 Abundance of smRNAs

Male smRNAs map to the AaegL5 reference genome assembly with a greater depth than female smRNAs across the three chromosomes, with some locations having particularly high coverage depth, which is possibly indicative of piRNA clusters. No clusters of high smRNA density appear to be associated with the M locus (Figure 4.13)

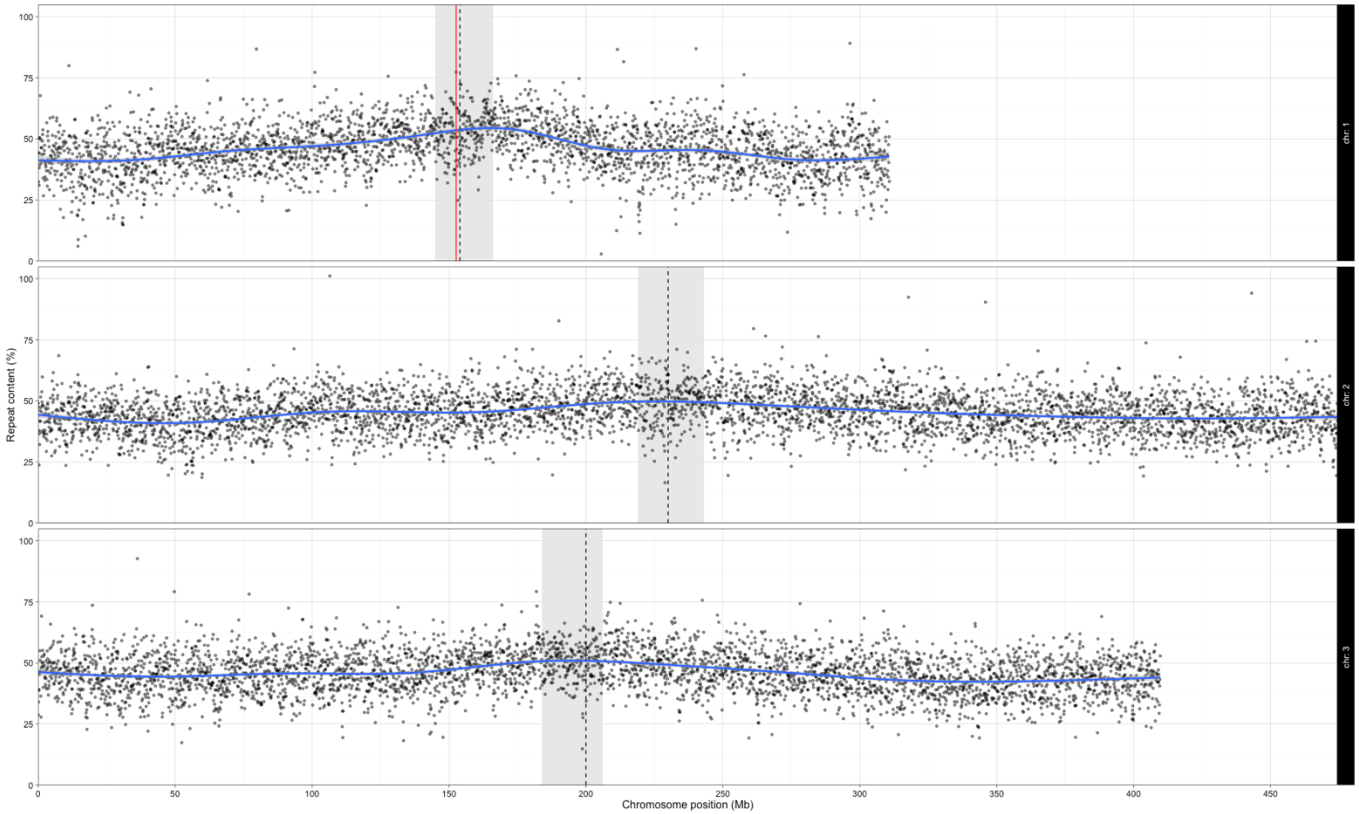


Figure 4.10 The percentage coverage of all classes of transposable elements across 100 kb bins throughout the AaegL5 Ae. aegypti *reference genome assembly. Each point represents a 100 kb bin. The blue line is the smoothed conditional mean. The red vertical line indicates the position of the M locus; the black dashed lines indicate the positions of the centromeres for each chromosome, and the grey vertical boxes show the positions of the pericentromeric cytogenetic bands. Chromosomes are different sizes, resulting in data shown only for part of the window for chr 1 and 3.*

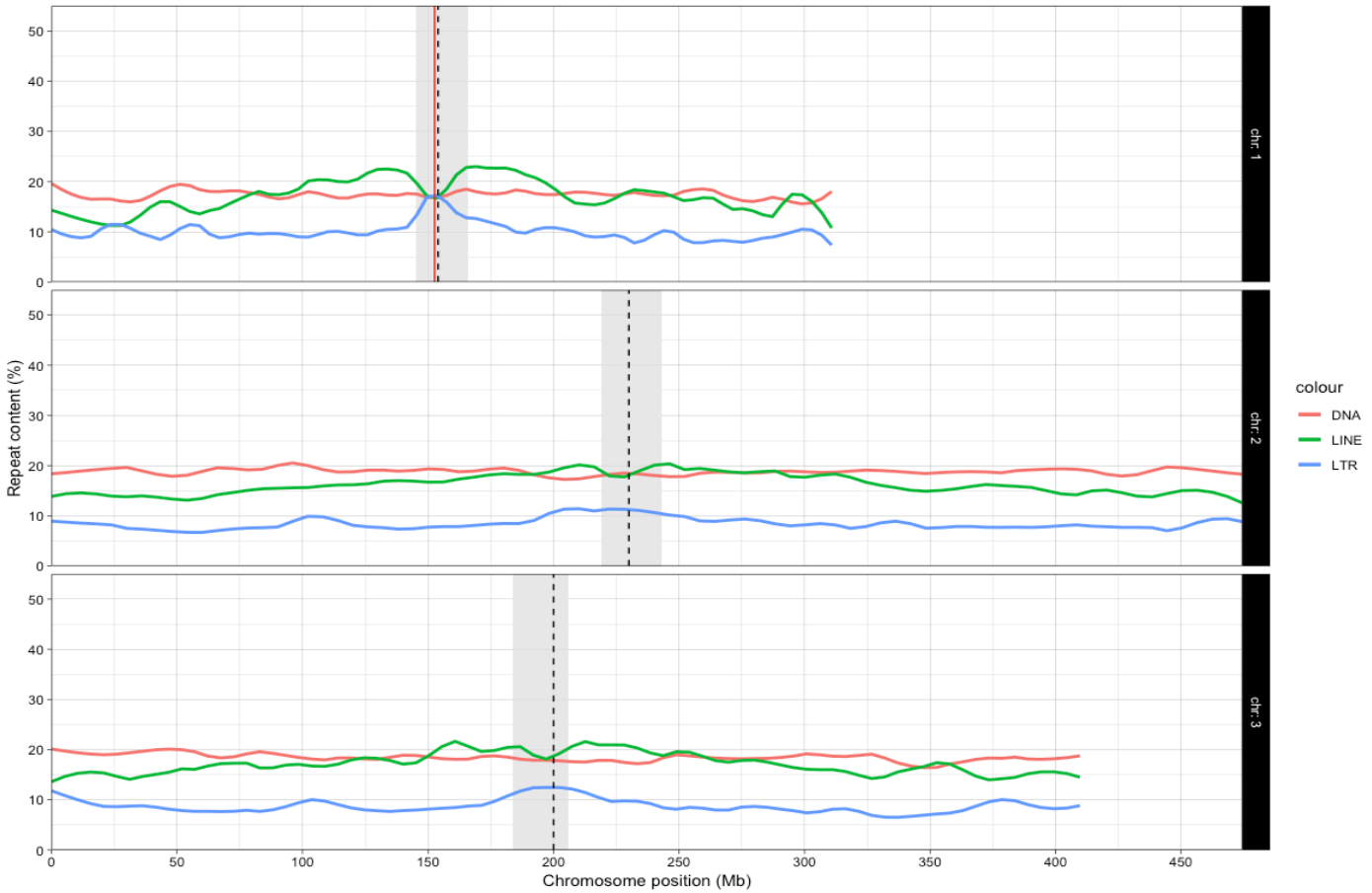


Figure 4.11 The percentage coverage different classes of transposable elements across 100 kb bins throughout the AaegL5 Ae. aegypti *reference genome assembly. Coloured lines are the smoothed conditional means (red: DNA transposons; green: LINEs; blue: LTRs). The red vertical line indicates the position of the M locus; the black dashed lines indicate the positions of the centromeres for each chromosome, and the grey vertical boxes show the positions of the pericentromeric cytogenetic bands. Chromosomes are different sizes, resulting in data shown only for part of the window for chr 1 and 3.*

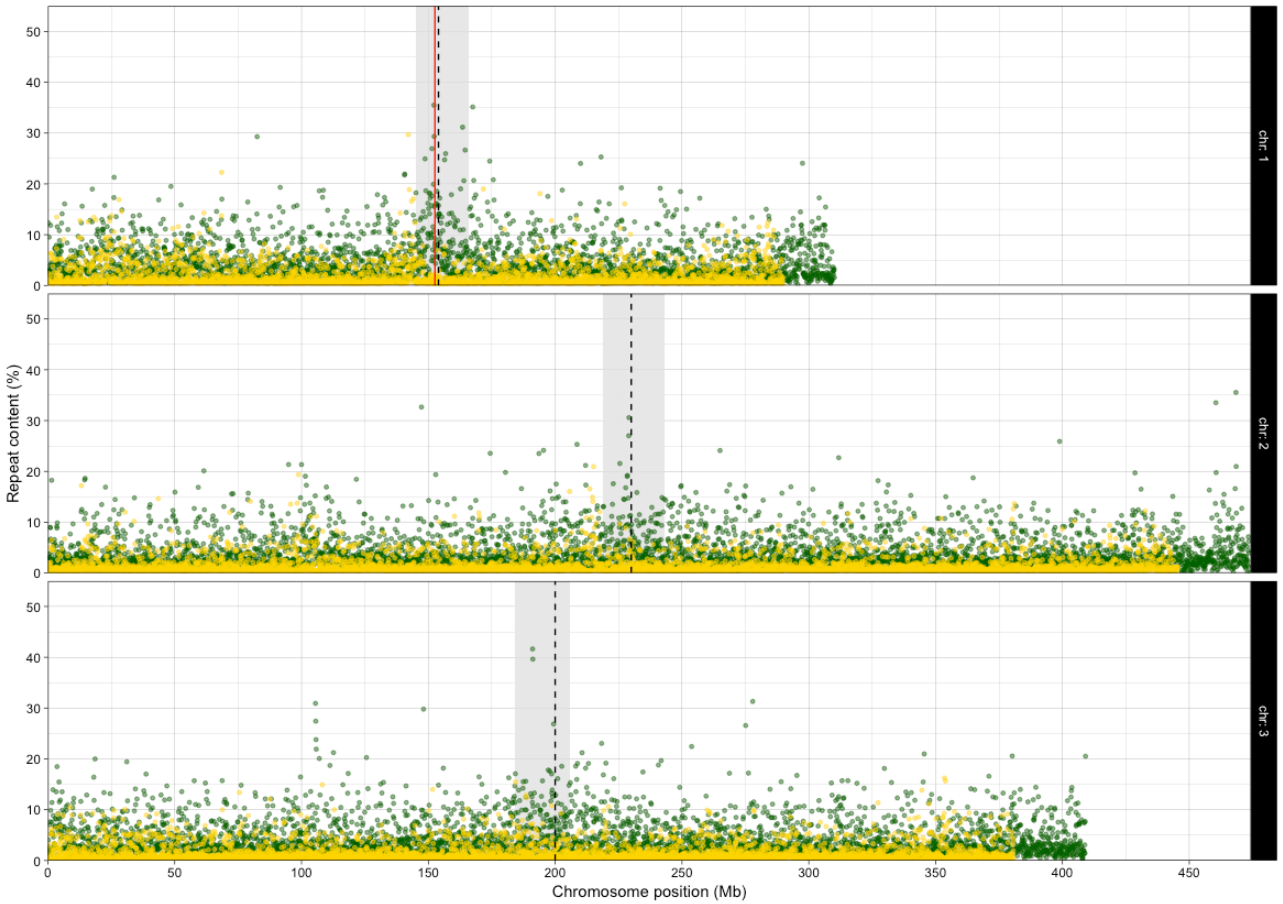


Figure 4.12 The percentage coverage of types transposable elements across 100 kb bins throughout the AaegL5 Ae. aegypti reference genome assembly (green: Gypsy; yellow: Copia). The red vertical line indicates the position of the M locus; the black dashed lines indicate the positions of the centromeres for each chromosome, and the grey vertical boxes show the positions of the pericentromeric cytogenetic bands. Chromosomes are different sizes, resulting in data shown only for part of the window for chr 1 and 3.

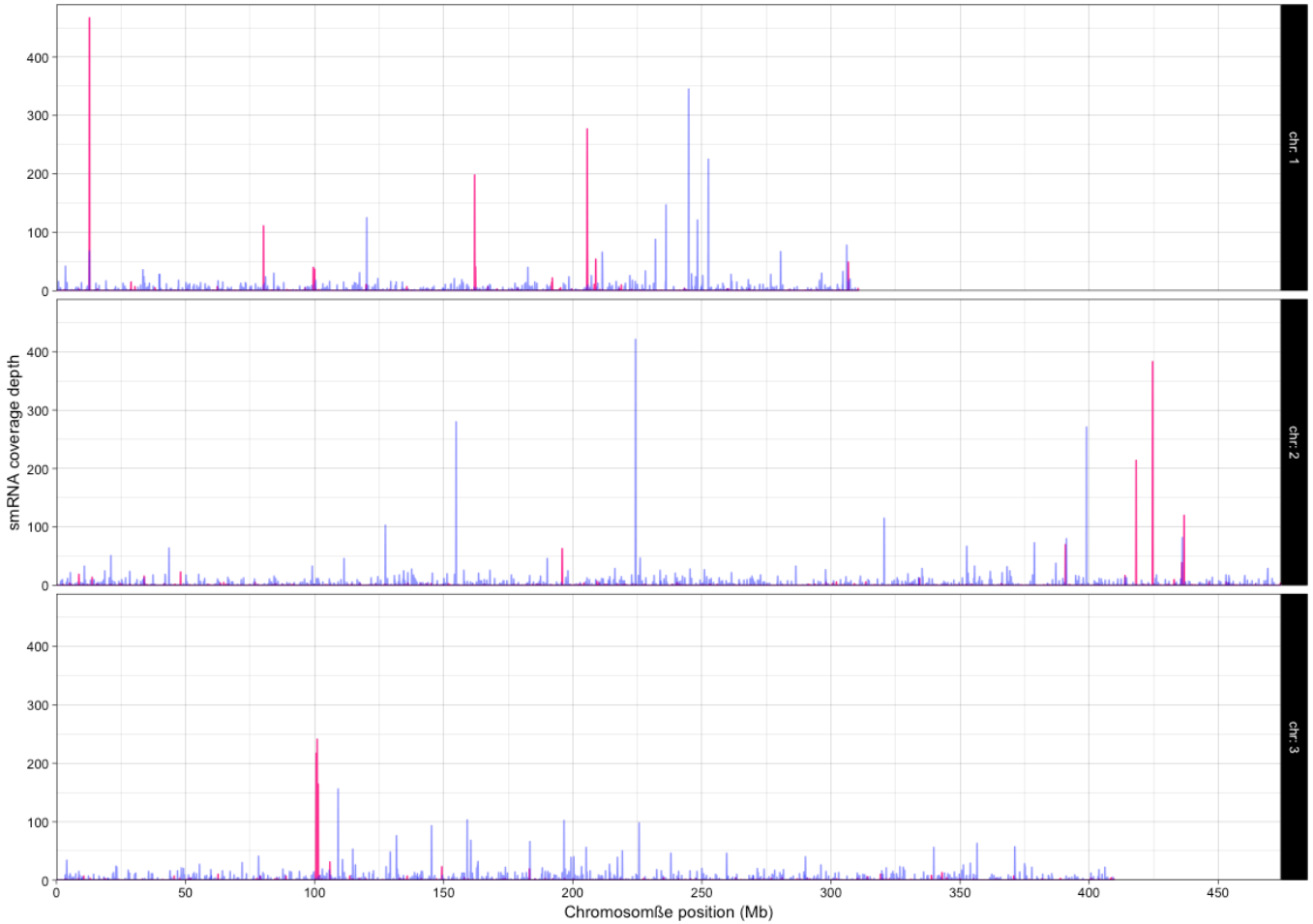


Figure 4.13 Depth of coverage of female (red) and male (blue) smRNA RNA-Seq reads across 100 kb bins throughout the AaegL5 Ae. aegypti *reference genome assembly. Chromosomes are different sizes, resulting in data shown only for part of the window for chr 1 and 3.*

4.4.6 Male-specific transcripts identified using a subtraction pipeline

Male RNA-Seq reads that did not map to the AaegL4 genome assembly were assembled into an initial transcriptome of 8,761 transcripts. The subtraction pipeline eliminated all but two transcripts that mapped to male, but not female, DNA, did not match any repeats in the *Ae. aegypti* library, and had at least twice the expression in male samples than female. The transcripts were queried against all known sequences with BLASTN. Both transcripts aligned with high identity to sequences in the AaegL5 genome assembly, although these were outside the M locus (Table 4.4): one matched a predicted non-coding RNA on chromosome 2 in the AaegL5 annotation (GenBank accession: XR_002500546.1), and the other aligned to part of chromosome 1 but is not associated with any annotated gene.

Table 4.4 Expression statistics and BLAST alignment details for the two putative male-specific transcripts identified from an Ae. aegypti de novo *transcriptome assembly using a subtraction pipeline.*

Candidate transcript	Female RPKM	Male RPKM	AaegL5 alignment start	AaegL5 alignment end	e-value	Gene hit
CL2Contig1	294,615	648,124	chr2:22,443,294	chr2:22,443,502	1e-84	LOC110676597
DN1118_c0_g3_i1	0	317,438	chr1:83,468,103	chr1:83,467,902	4e-86	–

4.4.7 Male-biased sequences in *Ae. albopictus*

The *Ae. albopictus* orthologues of *Nix* and *myo-sex* are present on separate contigs (Table 4.5); the *Nix*-containing contig, NW_017857498, shows greater male coverage than female across its entire length, although it does not contain a greater proportion of transposable elements than the genome average (Figure 4.14). In contrast, the *myo-sex*-containing contig, NW_017856377, does not exhibit male-specific coverage (Table 4.5). The differential coverage pipeline identified the *Nix*-containing contig as the 14th most male-biased contig in the genome, while the *myo-sex*-containing contig was not in the 200 most sex-differentiated contigs, indicating that there may be other sequences in the current genome assembly that form part of

the *Ae. albopictus* M locus. Comparison of the *Ae. albopictus Nix* contig with the *Nix* chromosome region of *Ae. aegypti* found the sequences exhibit low similarity (Supplementary Figure 3).

Table 4.5 Sex-specific coverage and repeat content statistics for two contigs containing the orthologues of the Ae. aegypti *M locus genes in the* Ae. albopictus *cell line genome assembly.*

Contig	Gene	Contig length (bp)	Female coverage breadth	Male coverage breadth	CQ	Repeat content (%)
NW_017857498	LOC109412105 (*Nix*)	970,929	0.929	0.998	0.780	52.6
NW_017856377	LOC109397226 (*myo-sex*)	3,758,540	0.895	0.893	0.980	54.7

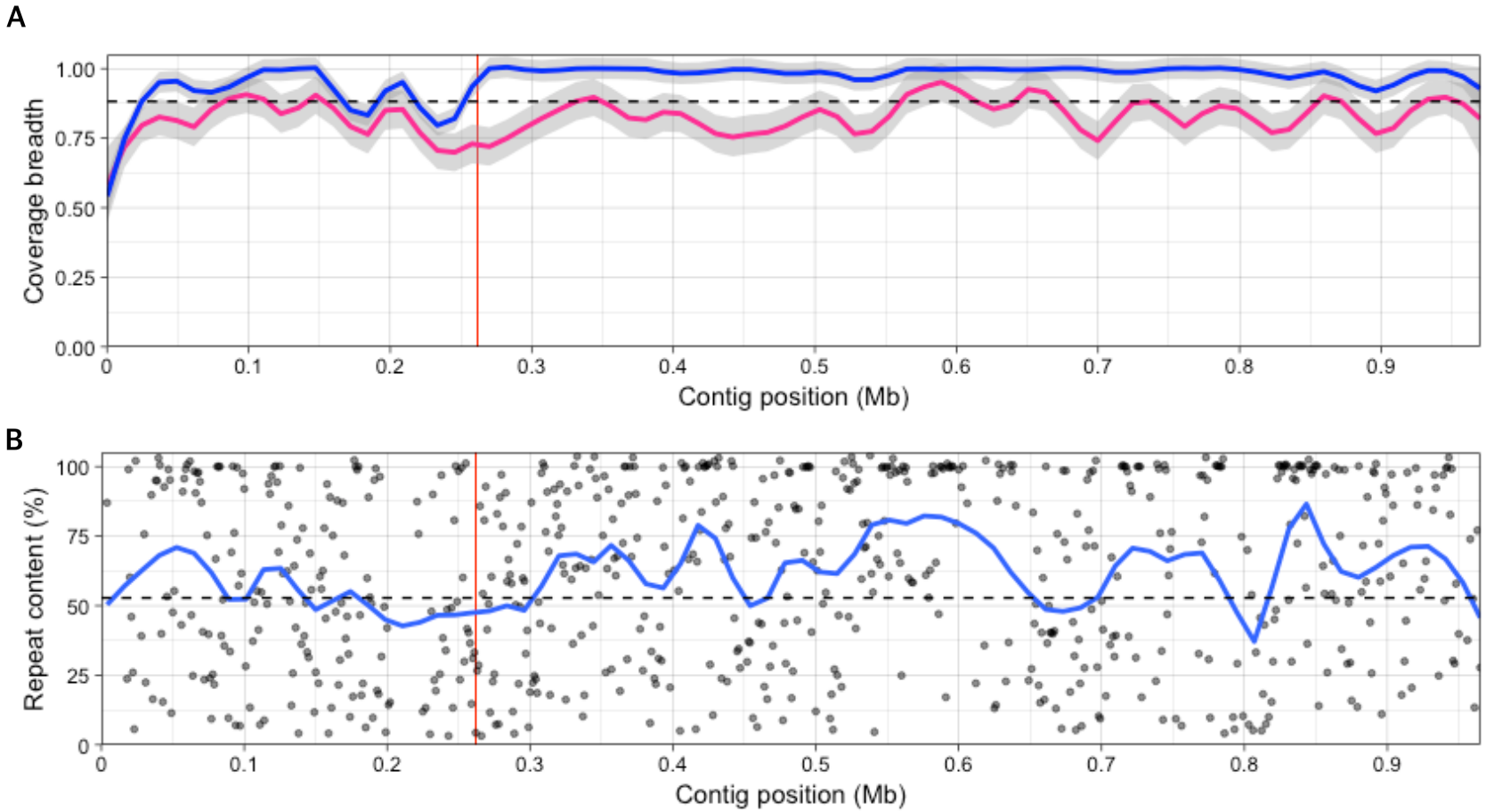


Figure 4.14 ***A*** *Breadth of coverage of female (red) and male (blue) genomic reads, with grey shading indicating 95% confidence intervals; and* ***B*** *percentage coverage of all classes of transposable elements; across 1 kb bins on a putative M locus contig (NW_017857498) in the* Ae. albopictus *cell line C6/36 genome assembly. The dotted horizontal lines represent the average coverage and the average repeat content across the entire genome in* ***A*** *and* ***B****, respectively. The red vertical line indicates the position of the* Nix *orthologue.*

4.5 Discussion

4.5.1 The M locus is contained within a wider sexually differentiated chromosomal region

The microscope images from the FISH experiments, using probes created from BAC sequences containing the male-specific genes *Nix* and *myo-sex* identified in Chapter 3, permitted the validation of the location of the M locus on physical chromosomes. Genomic analyses had located these genes in the AaegL5 assembly close to the centromere on the p arm of chromosome 1 (Matthews *et al.*, 2018), and the chromosomal FISH experiments provide strong additional evidence for this placement. Notably, the location of the M locus based on the full-sequence genes is contrary to previous studies, in which the same FISH protocol was used with the protein coding sequences of *Nix* and *myo-sex,* yet both genes hybridised at position 1q21 (Hall *et al.*, 2014; Hall *et al.*, 2015). The reasons for this discrepancy are uncertain, however the updated location is consistent with earlier genetic mapping studies, which used linkage analysis of genetic markers and Giemsa staining to determine that the M locus is very close to the centromere (Bhalla and Craig, 1970; Newton *et al.*, 1974; Motara and Rai, 1977; Newton *et al.*, 1978). Thus, these experiments most likely represent the most accurate assignment of the cytogenetic position of the M locus yet conducted.

Analysis of the coverage of DNA reads mapped to the AaegL5 assembly provided further validation of the maleness of the putative M locus, with substantially fewer female reads aligning to the hypothesised ~1.5 Mb region on chromosome 1 (Figure 4.3 and Figure 4.4). While it is unclear why the CGR data yielded incongruous results when applied to the CQ analysis, the dramatic decline in female coverage combined with the maintenance of high male coverage around *Nix* uncovered by the differential mapping analysis suggests that the M locus is similar in position and size in both strains. Alignment of male linked reads sequenced from the same "CGR" strain (AWT Family 2) also showed that variation between this strain and the AaegL5 genome sequenced strain (LVP_AGWG) was overwhelmingly present in a

continuous phased block over the M locus (Figure 4.8), further indicating that it is present as a single, male-specific haplotype.

A distinctive result of the differential mapping analysis is that female coverage tends to be lower than male coverage across a wider region of chromosome 1 than expected. This finding is consistent with recent studies of genetic polymorphism over this chromosome; for example, data from wild populations found that sex-specific genetic differentiation was generally higher on chromosome 1 than other chromosomes, especially in proximity to the M locus, while male-specific heterozygosity was enhanced around the M locus in a wild Thai (though not a Senegalese) strain (Campbell *et al.*, 2017). Another study found evidence of repressed recombination across ~40% of chromosome 1 using linkage mapping intercrossing experiments (Fontaine *et al.*, 2017). This same study also noted that analysis of RADseq markers revealed that chromosome 1 exhibited male-female differentiation between 148 – 211 Mb in two populations from Australia and Brazil, corresponding very precisely with the region of reduced female- relative to male-specific coverage in the CGR Illumina dataset (Figure 4.3 and Figure 4.4). This region appears to be larger in the Virginia Tech and Rockefeller data (Figure 4.6), which are both from closely-related Liverpool strains, but is consistent with observations of higher linkage disequilibrium in the Liverpool strain across a similar 100 Mb region (Fontaine *et al.*, 2017). The CGR data comes from a different laboratory strain, suggesting the size of the male region – and the extent of sex-specific differentiation – may vary across populations. Unfortunately, we cannot study this question in the Cambridge dataset, which comprises a large number of strains, as these are exome-captured data of known genes not including *Nix*, rather than whole-genome libraries, and contains few individuals known to be male (of the individuals that were sexed before sequencing, most were female; Crawford *et al.* 2017; see also Appendix 2.1).

Interestingly, all 35 contigs from the AaegL3 assembly with the greatest sex-biased coverage (d) aligned to a subset of this differentiated region somewhere between 25

– 50 Mb away from the M locus (Figure 4.6). While it was known that *Nix* was not present in earlier assemblies, this result suggests that no part of the M locus was incorporated into these assemblies. The location of this cluster of male sequences is relatively close to putative former genes that have been inactivated by mutation (null alleles) in an Australian population (Fontaine *et al.*, 2017). Null alleles are known to accumulate on Y chromosomes (Rice, 1987), raising the possibility that some of these sequences may form parts of genes that have become inactivated during the transition towards a Y chromosome; further work could investigate this.

4.5.2 The M locus is enriched for retrotransposons but not smRNA clusters

Analysis of the distribution of transposable elements found that they are prevalent around the centromeres of all three chromosomes, revealing that *Ae. aegypti* has heterochromatic centromeric regions that contain disproportionately higher levels of repetitive sequence, including TEs (Sun *et al.*, 2003; Severson *et al.*, 2004; Wong and Choo, 2004). However, the increased centromeric repeat density is particularly pronounced on chromosome 1, showing a shift from LINEs in favour of LTRs, especially Gypsy and Copia retrotransposons (Figure 4.10–Figure 4.12). These two elements are prominent TEs in *Drosophila*, with the ability to transmit horizontally in the genome and induce novel phenotypes (Corces and Geyer, 1991; Kim *et al.*, 1994; Jordan and McDonald, 1998), and they have also been characterised in detail in mosquitoes (Tu and Coates, 2004). Gypsy and Copia LTRs are also known to be prevalent in plant genomes, and in many cases have been found to be enriched in sex-differentiated regions, including non-recombining male-determining regions similar to the M locus (Liu *et al.*, 2004; Marais *et al.*, 2008; Kejnovsky *et al.*, 2009; Harkess *et al.*, 2017; Kudoh *et al.*, 2018). This process appears to be replicated around the *Ae. aegypti* M locus, potentially due to relaxed selection against TE insertions due to locally reduced recombination (Charlesworth *et al.*, 2005; Bachtrog, 2013).

In *Ae. aegypti*, as in *Drosophila* and other insects, non-coding small RNAs are involved in the regulation and silencing of retroviral sequences, to protect against the deleterious effects of both viruses and TEs (Malone and Hannon, 2009; Gammon and Mello, 2015; Lewis *et al.*, 2018). Some types of smRNAs like siRNAs and piRNAs are commonly derived from TEs (Biryukova and Ye, 2015; Miesen *et al.*, 2016), and it is known that the *Ae. aegypti* genome contains piRNA clusters associated with high TE content (Arensburger *et al.*, 2011; Palatini *et al.*, 2017; Whitfield *et al.*, 2017). It might therefore be expected that the accumulation of TEs around the M locus could be associated with a greater density of piRNA clusters in males, yet no increased coverage of male-specific smRNA data was observed at this locus (Figure 4.13). This is consistent with previous studies that have reported a lack of unique smRNAs associated with M-linked sequences (Adelman and Tu, 2016). However, the male and female data used in this chapter came from different strains, and future work could investigate whether a similar pattern is demonstrated in equivalent male and female smRNA datasets.

4.5.3 Future directions

The genomic analyses using the male AaegL5 reference assembly described in this chapter found strong evidence that the *Ae. aegypti* chromosome 1 displays characteristics of transitioning to a sex chromosome, such as diverged sequences on the male-limited (M) copy, as well as introgression of TEs, due to a reduction in recombination between the two chromosome pairs. However, gaps still remain in understanding the M locus, and further investigation – especially that using population genetics and comparative evolutionary data – could determine how its structure and evolution is shaped in this mosquito species.

One of the remaining mysteries is the nature of the boundaries between the M locus and the pseudoautosomal regions of the rest of chromosome 1. Although male-biased coverage extends over a large portion of the chromosome, implying that a large region may show divergence between the sex chromosome homologs due to an

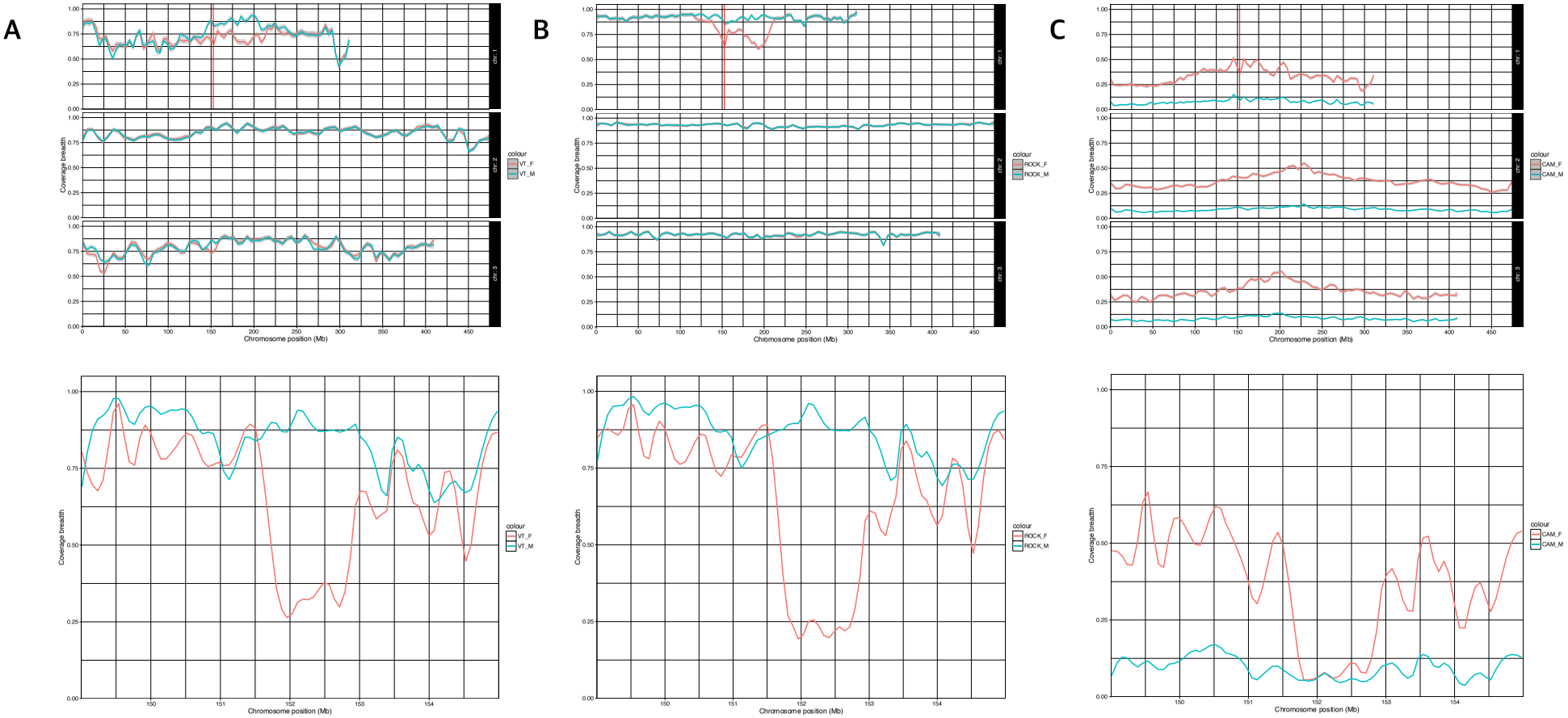
extensive region of reduced recombination, these regions have been shown to recombine, including – though rarely – the gene *myo-sex* located in the ~1.5 Mb putative M locus (Hall *et al.*, 2014). Therefore, it may be that *Nix* is the only truly male-limited gene that occurs in a region of completely suppressed recombination. *Nix* appears to be the only truly male-specific gene, although other non-coding transcripts such as those identified through the subtraction pipeline (Table 4.4) and alignment of candidate male-specific sequences (Figure 4.6) may have some current or former role. Previous studies posited that an inversion may be responsible for the breakdown in recombination between the M and m locus chromosome pairs and harbour the sex-determining locus (Hickey and Craig, 1966; Bhalla and Craig, 1970); however, the analysis of structural variation using 10x linked read data did not locate any inversion breakpoints in the vicinity of *Nix*. Finding the boundaries of the M locus may require population-level analysis; the improved AaegL5 genome assembly will allow analysis of genetic variation near the M locus, for instance using pooled-sample sequencing (Pool-seq; Kapun *et al.*, 2014; Schlötterer *et al.*, 2014). Although Pool-seq is useful for detecting population variation, it does not allow haplotypes to be discerned, and it may be more useful to sequence individual males and females from separate subpopulations. Such approaches could also be used to examine diversity of LTRs and investigate if they are fixed or variable across populations, or recently introduced and active or old and inactivated.

Another unsettled question is the age of the M locus and whether it is ancestral or recently derived from an alternative sex-determination system. Given that the M and m chromosomes behave somewhat like proto-Y and -X chromosomes on the trajectory towards differentiating into heteromorphic pairs, it might be expected that the M locus evolved relatively recently. However, the related mosquito species *Ae. albopictus* also has a *Nix* gene formed of two exons that shows 70% identity to its orthologue in *Ae. aegypti* (Miller *et al.*, 2018), and the contig on which it is located in the cell line assembly shows male-biased coverage similar to that in the AaegL5 M locus (Figure 4.14), although the surrounding regions show very little

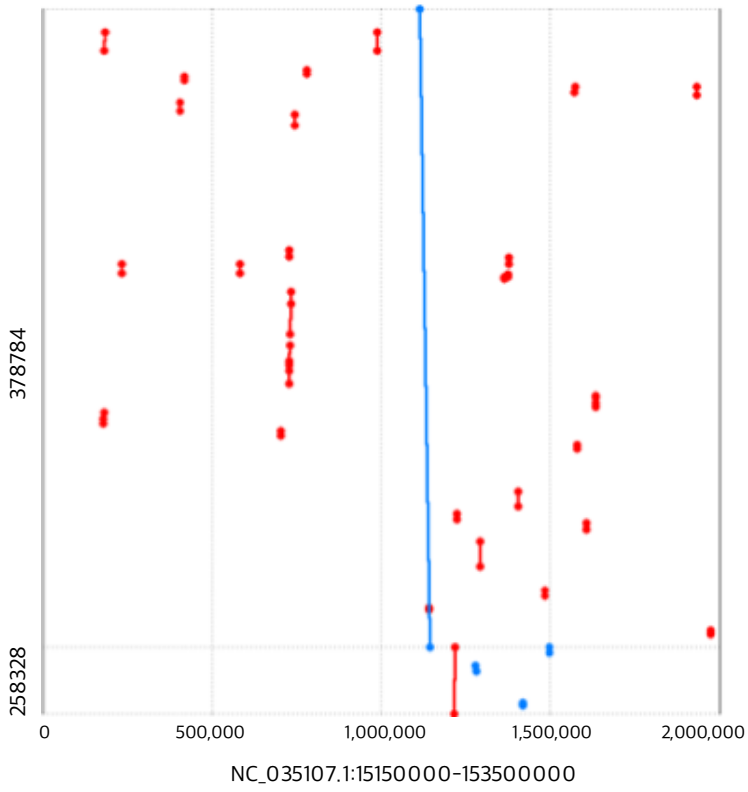
conserved sequence between the two species (Supplementary Figure 3). As insects have a high turnover of sex determination systems (Kaiser and Bachtrog, 2010; Vicoso and Bachtrog, 2015), it is plausible that *Nix* might have evolved a sex-determining function independently in both species, which are estimated to have diverged from a common ancestor approximately 71.4 Mya (Chen *et al.*, 2015), perhaps from a related alternative function. In-depth evolutionary comparisons could investigate this in more detail, and would benefit from high quality genomes for *Ae. albopictus* and other mosquito species in the *Aedes* genus, as well as other culicine genera with homomorphic sex chromosomes such as *Culex*.

The vastly improved completeness of the improved *Ae. aegypti* genome assembly, AaegL5, will facilitate research into these questions, and will make it a valuable resource for the mosquito community. This will help to expand knowledge of the structure, content, population variation and evolutionary history of the *Ae. aegypti* sex determination system, and may improve genetic vector control strategies that employ sex-specific targets.

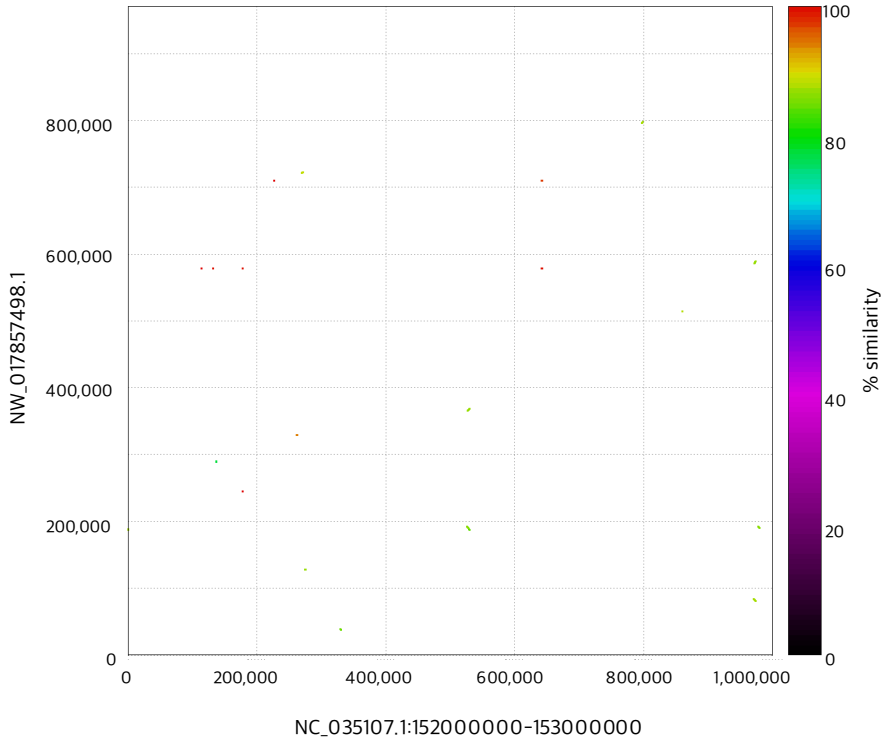
4.6 Supplementary data



Supplementary Figure 1 Breadth of coverage of female (red) and male (blue) genomic reads across 30 kb bins throughout the AaegL5 Ae. aegypti *reference genome assembly (top) and between 149–155 Mb on chromosome 1, encompassing the M locus (bottom) for different datasets:* ***A*** *Virginia Tech (LVP);* ***B*** *Rockefeller (AGWG_ LVP);* ***C*** *Cambridge (Appendix 2.1).*



Supplementary Figure 2 The two Nix*-containing contigs of the AWT Family 2 Supernova assembly (258328 and 378784;* y *axis) queried against the M locus region of chromosome 1 of the reference genome (NC_035107.1:151500000-153500000;* x *axis). Alignment and plot generated using MUMmer 4.0.0* (*Kurtz* et al.*, 2004*).



Supplementary Figure 3 The Nix*-containing contig of the* Ae. albopictus *C6/36 cell line genome (NW_ 017857498;* y *axis) queried against part of the M locus on chromosome 1 of the* Ae. aegypti *genome (NC_ 035107.1:152000000-153000000;* x *axis). Alignment and plot generated using MUMmer 4.0.0* (*Kurtz* et al.*, 2004*).

Supplementary Table 1 Top BLAST hits for 35 male-biased contigs from the AaegL3 reference assembly queried against the chromosome-length AaegL5 reference assembly.

Query_contig_ AaegL3	Query_ length	Subject_chr_ AaegL5	Subject_ start	Subject_ end	Alignment_ length
AAGE02035037.1	6261	1	186654721	186660984	6264
AAGE02035965.1	4651	1	170301555	170298998	2595
AAGE02035016.1	6297	1	188483148	188489444	6297
AAGE02035557.1	5426	1	183385439	183387180	1768
AAGE02036067.1	4251	1	195847559	195851811	4254
AAGE02035994.1	4546	1	191742908	191747452	4545
AAGE02034767.1	6814	1	187007233	187000541	6695
AAGE02033847.1	10980	1	193606092	193596688	9406
AAGE02034061.1	3832	1	191284416	191280590	3831
AAGE02034583.1	7261	1	195805187	195812443	7260
AAGE02034705.1	6954	1	192531394	192538342	6949
AAGE02034751.1	6843	1	196818091	196824914	6844
AAGE02034753.1	6842	1	183475980	183469140	6841
AAGE02034775.1	6804	1	183786271	183779466	6808
AAGE02034781.1	6794	1	200802480	200795689	6794
AAGE02034898.1	6548	1	187238879	187232333	6548
AAGE02034914.1	6515	1	195795966	195802480	6515
AAGE02034940.1	6461	1	200620267	200613804	6464
AAGE02034986.1	6342	1	195773557	195767216	6343
AAGE02034996.1	6331	1	186457019	186463342	6337
AAGE02035035.1	6264	1	195758569	195752303	6267
AAGE02035131.1	6099	1	191459309	191453212	6098
AAGE02035307.1	5808	1	191510768	191504956	5813
AAGE02035343.1	5751	1	191084521	191089727	5240
AAGE02035403.1	5641	1	197801893	197807530	5639
AAGE02035462.1	5563	1	183544671	183550233	5564
AAGE02035539.1	5454	1	177753519	177758970	5453
AAGE02035553.1	5428	1	182304699	182299278	5428
AAGE02035639.1	5277	1	175654209	175648933	5278
AAGE02035662.1	5226	1	193397812	193403036	5225
AAGE02035772.1	5039	1	195273640	195272558	1088
AAGE02035790.1	5012	1	196279657	196284667	5011
AAGE02035906.1	4795	1	195021683	195026479	4798
AAGE02036165.1	3746	1	187021770	187025509	3745
AAGE02036194.1	3234	1	187915575	187912344	3233

Chapter 5 General Discussion

5.1 Genomics of sex chromosome evolution in *Aedes aegypti*

This thesis aimed to elucidate the genetic mechanism of sex determination in the dengue vector mosquito *Ae. aegypti* and test potential synthetic biology techniques for reliably targeting transgenic constructs to either sex, in order to improve genetic techniques for controlling mosquito populations and limiting the spread of disease. At the beginning of the project, understanding of the genetic basis of sex in this mosquito species was quite limited. While some ingenious classical genetics studies in the mid-20th century determined the existence of a male determining factor on a non-recombining section of chromosome 1 (e.g. Craig *et al.*, 1960; Bhalla and Craig, 1970; Newton *et al.*, 1978), and later studies described the structure and function of *doublesex* (*dsx*), which triggers the downstream somatic sex determination cascade (e.g. Salvemini *et al.*, 2011), and investigated differences in male and female gene expression (e.g. Tomchaney *et al.*, 2014), the initial genetic switch evaded discovery due to its location within a genomic "black box": the M locus, absent from early genome assemblies.

Since the project began in 2014, knowledge of the M locus has increased, both from the work in this thesis and work conducted by other researchers, and informed by genomics at each stage: In 2015, researchers identified the M locus gene *Nix* using whole genome Illumina data applied to the Chromosome Quotient (CQ) method, and collated strong evidence that it acts as the primary sex determination signal, initiating male development by modulating the splicing of *dsx* in some way (Hall *et al.*, 2015). Later, the full sequence of *Nix* and its genomic context was ascertained using PacBio sequencing, describing its exceptionally large intron and the accrual of transposons (Turner *et al.* 2018; Chapter 3). After this, PacBio sequencing and Hi-C on male DNA was used to assemble an improved genome assembly anchored to chromosomes and containing nearly the full sequence of the M locus (Matthews *et al.*, 2018; Chapter 4). Chapter 4 also investigated the wider genomic architecture of the M chromosome using 10x and Illumina data.

These findings illustrate the importance of genomics in understanding sex determination and sex chromosome biology. Early genomics studies were largely confined to model species, such as *Drosophila melanogaster* in insects (Adams *et al.*, 2000), due to the high cost of sequencing DNA. Although *D. melanogaster* is a well-studied model organism, its ancestor is estimated to have separated from the common ancestor with *Aedes* approximately 260 Mya (Chen *et al.*, 2015), and its genome has many divergent characteristics (Wiegmann and Richards, 2018). Crucially, the sex chromosomes of the two insects are dissimilar, and sex determination systems are known to evolve rapidly in flies and other insects (Kaiser and Bachtrog, 2010; Bopp *et al.*, 2014; Vicoso and Bachtrog, 2015).

Given the reduced costs of sequencing in recent years, this further underscores the value of pursuing genomic information for a range of species. Genome sequencing projects and subsequent incremental improvements to reference genomes are becoming commonplace in mosquitoes (e.g. Holt *et al.*, 2002; Arensburger *et al.*, 2010; Neafsey *et al.*, 2015; Miller *et al.*, 2018). In addition to wider taxonomic sampling, the utility of the improved *Ae. aegypti* reference AaegL5 demonstrates the benefits of high quality, single haplotype chromosome-length genome assemblies from a single sex for studying mosquito sex chromosomes. Despite the large budgets associated with such endeavours, which require a variety of expensive sequencing data to put together, diminishing costs of sequencing and computational technologies may mean that such assemblies can be generated for a greater number of species in the near future.

The research presented in this thesis shows that genomic analyses can uncover insights into the process of sex chromosome evolution in mosquitoes. Analysis of the *Nix* region in Chapter 3, followed by genome-wide analysis in Chapter 4, showed that the smallest pair of chromosomes (chromosome 1), on one pair of which the M locus resides in males, appear to be evolving towards differentiated sex chromosomes. Transposable elements (TEs) have built up both at the site of the M locus gene *Nix* – including within its intron, expanding its length greatly despite the expected selective advantage for a short, rapidly transcribed intron expressed in

early embryonic development – and within the wider chromosomal region in the vicinity of the M locus. Here, the density of retrotransposons is especially pronounced in comparison to the corresponding sections of the other two chromosomes, particularly the two LTR classes Gypsy and Copia, which are sometimes associated with Y chromosomes in *Drosophila* (Chang and Larracuente, 2018; Mahajan *et al.*, 2018), *Anopheles* (Hall *et al.*, 2016), and some plants (e.g. Liu *et al.*, 2004; Kudoh *et al.*, 2018). This dynamic is frequently representative of young sex chromosomes due to the relaxed selection pressure against insertion of TEs resulting from the suppression of recombination between the homologous pairs (Bachtrog, 2013; see also Figure 1.2).

Further evidence of an ongoing transition to differentiated sex chromosomes is the male-specific coverage spanning a large pseudoautosomal section of the male copy of chromosome 1. While the ratio of female to male coverage is lowest at *Nix* and surrounding 1.5 Mb "proper" M locus, male bias spans a wider sequence than just this locus. Previously it was thought that the chromosome pairs were fully undifferentiated outside of the M locus, but more recent findings comparing sex-specific variation on this chromosome report very consistent evidence of male divergence (Campbell *et al.*, 2017; Fontaine *et al.*, 2017), suggesting that recombination is reduced across a larger part chromosome, possibly as a precursor to near-total suppression.

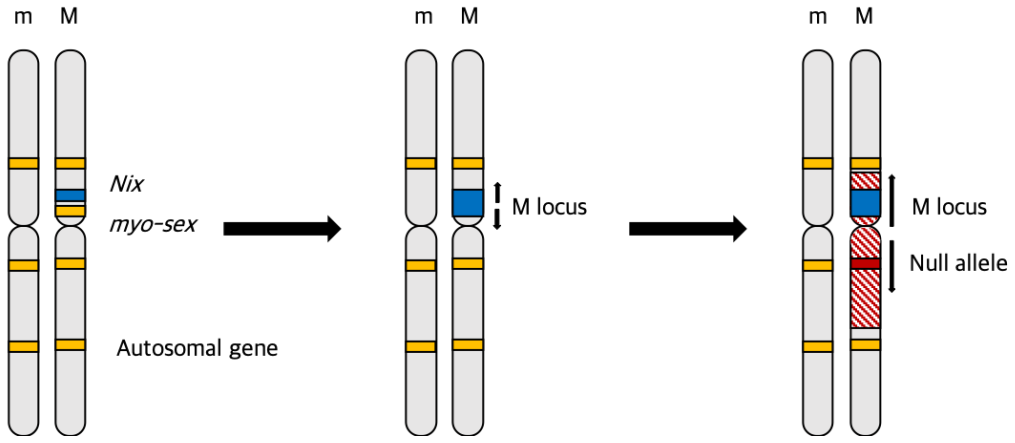


Figure 5.1 A hypothetical schematic for the evolution of the Ae. aegypti *M locus. Initially, the sex determination gene* Nix *(blue) arises on the autosome. Next, the suppression of recombination at* Nix *is*

favoured to limit its inheritance to the male sex, and an M locus (blue) encompassing myo-sex *emerges. Eventually there is reduction in recombination over a region larger than the M locus (red shading), and some alleles on the M chromosome become inactivated (red).*

However, this explanation is made more complex by the presence of an orthologue of *Nix* in *Ae. albopictus* (Chen *et al.*, 2015; Miller *et al.*, 2018), and the predominance of undifferentiated sex chromosomes in *Aedes* and *Culex*. It is hypothesised that homomorphic chromosomes were the ancestral state in the common ancestor of anopheline and culicine mosquitoes, and *Anopheles* subsequently developed a Y chromosome (Toups and Hahn, 2010). Unless the *Nix* evolved convergently in *Ae. aegypti* and *Ae.* albopictus, the probable shared ancestry between the sex determination genes in these species suggests that homomorphic chromosomes have been maintained in *Aedes,* which would make them many millions of years old. This raises the question of why *Ae. aegypti* has not transitioned to an XY arrangement in this time, given that it exhibits many genetic characteristics of undergoing this transition. Male genomic information from other mosquito species could help to answer this question, as well as comparisons to other organisms with similar chromosome arrangements. For instance, in ratite birds like emus, sex-biased gene expression helps to alleviate sexual antagonism between genes with differential fitness effects between sexes, maintaining recombination and homomorphic sex chromosomes (Vicoso *et al.*, 2013). In the brown alga *Ectocarpus* with haploid UV sex determination, the ~1 Mb sex-determining region is estimated to be over 100 Mya, and accumulation of TEs and sequence degeneration is very limited outside of the region (Ahmed *et al.*, 2014). Future work could examine the similarities between these stable non-recombining systems and *Ae. aegypti*. Additionally, population genomics data could be used to analyse potential variation in the extent of the M locus and the wider non-recombining chromosomal region, and identify features responsible for the suppression of recombination. Such analyses could generate insights into the factors that lead to either the stability and preservation of homomorphy or the transition to XY chromosomes in mosquitoes.

Close and distantly related subpopulations could be compared to attempt to discern the evolutionary trajectory of the M and m chromosomes. For instance, the outgroup *Ae. mascarensis*, found in Mauritius, is the closest existing relative to *Ae. aegypti* (Gloria-Soria *et al.*, 2016), and the two can form hybrids. However, an intersex phenotype is observed when male hybrids are backcrossed to *Ae. aegypti* (Motara and Rai, 1977), suggesting that the difference between the chromosomes in these species could present a useful example to investigate the evolution of sex determination.

5.2 Future directions for the genetic control of mosquitoes

Besides their interest from an evolutionary point of view, the results presented in this thesis may have some relevance to the development of genetic techniques for mosquito control. *Ae. aegypti* is the primary vector of dengue, chikungunya, and Zika, which represent a significant disease burden in the tropics. Warming global temperatures and urbanisation, coupled with rising resistance to insecticides, is expected to hasten the expansion of *Ae. aegypti* and other vector species such as *Ae. albopictus* (Kraemer *et al.*, 2015). Responding to this threat will require careful consideration of the available strategies for disease control, and may integrate genetic technologies for controlling mosquito populations (Alphey, 2014; World Health Organization, 2016). Research into the genomics of sex determination in mosquitoes is therefore important for improving these technologies, for example through allowing genetic sexing or sex-specific targeting of genetically encoded effects (Adelman and Tu, 2016). Genetic sexing strains are possible without precise knowledge of sex chromosome content, for instance by utilising the gene *dsx* (Fu *et al.*, 2007; Totten *et al.*, 2013; Hoang *et al.*, 2016), however it makes them more predictable and reliable if more is known about the genetic basis of sex determination, and precise editing of sex chromosomes is still a promising method for introducing sex-specific effects.

CRISPR/Cas9 will likely be an important tool in engineering these effects. Chapter 2 describes the attempts to modify the M locus by precisely integrating a fluorescence gene, with the intention of later replicating this with other constructs to institute sex-specific effects. Although these attempts were unsuccessful, the generation of a transgenic mosquito line expressing Cas9 in the germline, which had not been achieved in the published literature at the time, may facilitate future CRISPR engineering of *Ae. aegypti.*

Similarly to the amount of information available about the *Ae. aegypti* M locus, the capabilities of CRISPR-based gene editing have rapidly improved since they first started to be widely tested shortly before this project began. Recent studies have successfully improved the specificity of gene editing, reducing off-target effects, and overcome restrictions on the range of targetable sequences (Akcakaya *et al.*, 2018; Nishimasu *et al.*, 2018). Importantly, Buchman and Akbari (2018) report successfully integrating constructs on the repetitive *D. melanogaster* Y chromosome, and it would be interesting to replicate this technique at the *Ae. aegypti* M locus.

The limitations of the Chapter 2 study further illuminate the importance of attaining accurate genomics data when genetically engineering mosquitoes and informing disease control. Though little was known about the *Ae. aegypti* sex-determining region at the time, the initial targets for transformation, identified as unplaced contigs in the fragmented AaegL3 genome assembly, are now known to be quite distant from the M locus. The improved reference assembly, AaegL5, will allow more precise construction of targets for genetic modification, and may facilitate more effective vector control technologies.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Venter, J. C., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Adelman, Z. N. and Tu, Z. (2016). Control of mosquito-borne infectious diseases: sex and gene drive. *Trends Parasitol.* **32**, 219–229.

Adelman, Z. N., Jasinskiene, N., Onal, S., Juhn, J., Ashikyan, A., Salampessy, M., MacCauley, T. and James, A. A. (2007). *nanos* gene control DNA mediates developmentally regulated transposition in the yellow fever mosquito *Aedes aegypti*. *Proc. Natl. Acad. Sci.* **104**, 9970–9975.

Ahmed, S., Cock, J. M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A. F., Dittami, S. M., Corre, E., Valero, M., Aury, J. M., Roze, D., Van De Peer, Y., Bothwell, J., Marais, G. A. B. and Coelho, S. M. (2014). A haploid system of sex determination in the brown alga *Ectocarpus* sp. *Curr. Biol.* **24**, 1945–1957.

Akbari, O. S., Matzen, K. D., Marshall, J. M., Huang, H., Ward, C. M. and Hay, B. A. (2013). A synthetic gene drive system for local, reversible modification and suppression of insect populations. *Curr. Biol.* **23**, 671–677.

Akcakaya, P., Bobbin, M. L., Guo, J. A., Malagon-Lopez, J., Clement, K., Garcia, S. P., Fellows, M. D., Porritt, M. J., Firth, M. A., Carreras, A., Baccega, T., Seeliger, F., Bjursell, M., Tsai, S. Q., Nguyen, N. T., Nitsch, R., Mayr, L. M., Pinello, L., Bohlooly-Y, M., Aryee, M. J., Maresca, M. and Joung, J. K. (2018). In vivo CRISPR editing with no detectable genome-wide off-target mutations. *Nature* **561**, 416–419.

Aliota, M. T., Peinado, S. A., Velez, I. D. and Osorio, J. E. (2016). The wMel strain of *Wolbachia* reduces transmission of Zika virus by *Aedes aegypti*. *Sci. Rep.* **6**, 28792.

Alphey, L. (2014). Genetic control of mosquitoes. *Annu. Rev. Entomol.* **59**, 205–224.

Alphey, L. (2016). Can CRISPR-Cas9 gene drives curb malaria? *Nat. Biotechnol.* **34**, 149–150.

Alphey, L., McKemey, A., Nimmo, D., Neira Oviedo, M., Lacroix, R., Matzen, K. and Beech, C. (2013). Genetic control of *Aedes* mosquitoes. *Pathog. Glob. Health* **107**, 170–179.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10.

Arensburger, P., Megy, K., Waterhouse, R. M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F., Campbell, C. L., Campbell, K. S., Atkinson, P. W., et al. (2010). Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* **330**, 86–88.

Arensburger, P., Hice, R. H., Wright, J. A., Craig, N. L. and Atkinson, P. W. (2011). The mosquito *Aedes aegypti* has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. *BMC Genomics* **12**, 606.

Artieri, C. G. and Fraser, H. B. (2014). Transcript length mediates developmental timing of gene expression across *Drosophila*. *Mol. Biol. Evol.* **31**, 2879–2889.

Aryan, A., Anderson, M. A. E., Myles, K. M. and Adelman, Z. N. (2013). TALEN-based gene disruption in the dengue vector *Aedes aegypti*. *PLoS One* **8**, e60082.

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124.

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T. L., Hahn, M. W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., Vamosi, J. C., Mank, J. E., Peichel, C. L., Ashman, T. L., Blackmon, H., Goldberg, E. E., Hahn, M. W., Kirkpatrick, M., Kitano, J., Mayrose, I., Ming, R., Pennell, M. W., Perrin, N., Valenzuela, N. and

Vamosi, J. C. (2014). Sex Determination: Why So Many Ways of Doing It? *PLoS Biol.* **12**, e1001899.

Bailly-Bechet, M., Haudry, A. and Lerat, E. (2014). "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. *Mob. DNA* **5**, 13.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. and Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477.

Barrangou, R. and Doudna, J. A. (2016). Applications of CRISPR technologies in research and beyond. *Nat. Biotechnol.* **34**, 933–941.

Bassett, A. R. and Liu, J. L. (2014). CRISPR/Cas9 and genome editing in *Drosophila*. *J. Genet. Genomics* **41**, 7–19.

Bassett, A. R., Tibbit, C., Ponting, C. P. and Liu, J. L. (2013). Highly efficient targeted mutagenesis of *Drosophila* with the CRISPR/Cas9 system. *Cell Rep.* **4**, 220–228.

Basu, S., Aryan, A., Overcash, J. M., Samuel, G. H., Anderson, M. A. E., Dahlem, T. J., Myles, K. M. and Adelman, Z. N. (2015). Silencing of end-joining repair for efficient site-specific gene insertion after TALEN/CRISPR mutagenesis in *Aedes aegypti*. *Proc. Natl. Acad. Sci.* **112**, 4038–4043.

Benedict, M. Q. and Robinson, A. S. (2003). The first releases of transgenic mosquitoes: An argument for the sterile insect technique. *Trends Parasitol.* **19**, 349–355.

Bergkessel, M. and Guthrie, C. (2013). Colony PCR. *Methods Enzymol.* **529**, 299–309.

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M. and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630.

Bernardini, F., Galizi, R., Menichelli, M., Papathanos, P.-A., Dritsou, V., Marois, E., Crisanti, A. and Windbichler, N. (2014). Site-specific genetic

engineering of the *Anopheles gambiae* Y chromosome. *Proc. Natl. Acad. Sci.* **111**, 7600–7605.

Bernardini, F., Galizi, R., Wunderlich, M., Taxiarchi, C., Kranjc, N., Kyrou, K., Hammond, A., Nolan, T., Lawniczak, M. N. K., Papathanos, P. A., Crisanti, A. and Windbichler, N. (2017). Cross-species Y chromosome function between malaria vectors of the *Anopheles gambiae* species complex. *Genetics* **207**, 729–740.

Beukeboom, L. W. and Perrin, N. (2014). *The Evolution of Sex Determination*. 1st ed. Oxford University Press.

Bhalla, S. C. and Craig, G. B. (1970). Linkage analysis of chromosome 1 of *Aedes aegypti*. *Can. J. Genet. Cytol.* **12**, 425–435.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F., George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J. and Hay, S. I. (2013). The global distribution and burden of dengue. *Nature* **496**, 504–507.

Biedler, J. K. and Tu, Z. (2016). Sex Determination in Mosquitoes. *Adv. In Insect Phys.* **51**, 37–66.

Biedler, J. K., Hu, W., Tae, H. and Tu, Z. (2012). Identification of early zygotic genes in the yellow fever mosquito *Aedes aegypti* and discovery of a motif involved in early zygotic genome activation. *PLoS One* **7**, e33933.

Bier, E., Harrison, M. M., O'connor-Giles, K. M. and Wildonger, J. (2018). Advances in engineering the fly genome with the CRISPR-Cas system. *Genetics* **208**, 1–18.

Biryukova, I. and Ye, T. (2015). Endogenous siRNAs and piRNAs derived from transposable elements and genes in the malaria vector mosquito *Anopheles gambiae*. *BMC Genomics* **16**, 1–17.

Bopp, D., Saccone, G. and Beye, M. (2014). Sex determination in insects: Variations on a common theme. *Sex. Dev.* **8**, 20–28.

Bouzidi, M. F., Franchel, J., Tao, Q., Stormo, K., Mraz, A., Nicolas, P. and

Mouzeyar, S. (2006). A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions. *Theor. Appl. Genet.* **113**, 81–89.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Korf, I. F., et al. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 2010–2047.

Brady, O. J., Gething, P. W., Bhatt, S., Messina, J. P., Brownstein, J. S., Hoen, A. G., Moyes, C. L., Farlow, A. W., Scott, T. W. and Hay, S. I. (2012). Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLoS Negl. Trop. Dis.* **6**, e1760.

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527.

Broad Institute Picard. https://broadinstitute.github.io/picard/.

Buchman, A. and Akbari, O. S. (2018). Site-specific transgenesis of the *D. melanogaster* Y-chromosome using CRISPR/Cas9. *bioRxiv* 310318.

Burt, A. (2003). Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proc. R. Soc. London B Biol. Sci.* **270**, 921–928.

Bushnell, B. BBMap. https://sourceforge.net/projects/bbmap/.

Campbell, C. L., Dickson, L. B., Lozano-Fuentes, S., Juneja, P., Jiggins, F. M. and Black, W. C. (2017). Alternative patterns of sex chromosome differentiation in *Aedes aegypti* (L). *BMC Genomics* **18**, 943.

Carvalho, A. and Clark, A. (2013). Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res.* **23**, 1894–1907.

Carvalho, A. B., Dobo, B. A., Vibranovski, M. D. and Clark, A. G. (2001). Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **98**, 13225–13230.

Carvalho, D. O., Nimmo, D., Naish, N., McKemey, A. R., Gray, P., Wilke, A. B. B., Marrelli, M. T., Virginio, J. F., Alphey, L. and Capurro, M. L. (2014).

Mass production of genetically modified *Aedes aegypti* for field releases in Brazil. *J. Vis. Exp.* e3579.

Carvalho, D. O., McKemey, A. R., Garziera, L., Lacroix, R., Donnelly, C. a., Alphey, L., Malavasi, A. and Capurro, M. L. (2015). Suppression of a field population of *Aedes aegypti* in Brazil by sustained release of transgenic male mosquitoes. *PLoS Negl. Trop. Dis.* **9**, e0003864.

Catteruccia, F., Benton, J. P. and Crisanti, A. (2005). An *Anopheles* transgenic sexing strain for vector control. *Nat. Biotechnol.* **23**, 1414–1417.

Chaisson, M. J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics* **13**, 238.

Chang, C. and Larracuente, A. M. (2018). Heterochromatin-enriched assemblies reveal the sequence and organization of the *Drosophila melanogaster* Y chromosome. *Genetics* (Early online).

Charlesworth, B. (1991). Evolution of sex chromosomes. *Science* **251**, 1030–1033.

Charlesworth, B. (1996). The evolution of chromosomal sex determination and dosage compensation. *Curr. Biol.* **6**, 149–162.

Charlesworth, B. and Charlesworth, D. (2000). The degeneration of Y chromosomes. *Philos. Trans. R. Soc. London Ser. B-Biological Sci.* **355**, 1563–1572.

Charlesworth, D. and Mank, J. E. (2010). The birds and the bees and the flowers and the trees: Lessons from genetic mapping of sex determination in plants and animals. *Genetics* **186**, 9–31.

Charlesworth, D., Charlesworth, B. and Marais, G. (2005). Steps in the evolution of heteromorphic sex chromosomes. *Heredity (Edinb).* **95**, 118–128.

Chen, X.-G., Jiang, X., Gu, J., Xu, M., Wu, Y., Deng, Y., Zhang, C., Bonizzoni, M., Dermauw, W., Vontas, J., Armbruster, P., Huang, X., Yang, Y., Zhang, H., He, W., Peng, H., Liu, Y., Wu, K., Chen, J., Lirakis, M., Topalis, P., Van Leeuwen, T., Hall, A. B., Jiang, X., Thorpe, C., Mueller, R. L., Sun, C., Waterhouse, R. M., Yan, G., Tu, Z. J., Fang, X. and James, A. A.

(2015). Genome sequence of the Asian Tiger mosquito, *Aedes albopictus* , reveals insights into its biology, genetics, and evolution. *Proc. Natl. Acad. Sci.* 201516410.

Chen, J. S., Dagdas, Y. S., Kleinstiver, B. P., Welch, M. M., Sousa, A. A., Harrington, L. B., Sternberg, S. H., Joung, J. K., Yildiz, A. and Doudna, J. A. (2017). Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W. and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R. and Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054.

Clements, A. N. (1992). *The Biology of Mosquitoes*. London: Chapman & Hall.

Coates, C. J., Jasinskiene, N., Miyashiro, L. and James, A. A. (1998). *Mariner* transposition and transformation of the yellow fever mosquito, *Aedes aegypti*. *Proc. Natl. Acad. Sci.* **95**, 3748–3751.

Coelho, S. M., Gueno, J., Lipinska, A. P., Cock, J. M. and Umen, J. G. (2018). UV Chromosomes and Haploid Sexual Systems. *Trends Plant Sci.* **23**, 794–807.

Corces, V. G. and Geyer, P. K. (1991). Interactions of retrotransposons with the host genome: the case of the gypsy element of *Drosophila*. *Trends Genet.* **7**, 86–90.

Cortez, D., Marin, R., Toledo-Flores, D., Froidevaux, L., Liechti, A., Waters, P. D., Grützner, F. and Kaessmann, H. (2014). Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493.

Craig, G. B., Hickey, W. A. and Vandehey, R. C. (1960). An inherited male-producing factor in *Aedes aegypti*. *Science* **132**, 1887–1889.

Crawford, J., Alves, J., Palmer, W., Day, J., Sylla, M., Ramasamy, R., Surendran, S., Black, W. I., Pain, A. and Jiggins, F. (2017). Population genomics reveals that an anthropophilic population of *Aedes aegypti* mosquitoes in West Africa recently gave rise to American and Asian populations of this major disease vector. *BMC Biol.* **15**, 16.

Criscione, F., O'Brochta, D. A. and Reid, W. (2015). Genetic technologies for disease vectors. *Curr. Opin. Insect Sci.* **10**, 90–97.

Criscione, F., Qi, Y. and Tu, Z. (2016). GUY1 confers complete female lethality and is a strong candidate for a male- determining factor in *Anopheles stephensi*. *Elife* **5**, e19281.

Cui, Y., Sun, J. L. and Yu, L. L. (2017). Application of the CRISPR gene-editing technique in insect functional genome studies – a review. *Entomol. Exp. Appl.* **162**, 124–132.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M. and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510.

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., Therkildsen, N. O., Morikawa, M. and Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: An introduction to high-throughput sequencing data analysis. *Mol. Ecol. Resour.* **12**, 1058–1067.

Dong, S., Lin, J., Held, N. L., Clem, R. J., Passarelli, A. L. and Franz, A. W. E. (2015). Heritable CRISPR/Cas9-mediated genome editing in the yellow fever mosquito, *Aedes aegypti*. *PLoS One* **10**, e0122353.

Dong, Z. Q., Chen, T. T., Zhang, J., Hu, N., Cao, M. Y., Dong, F. F., Jiang, Y. M., Chen, P., Lu, C. and Pan, M. H. (2016). Establishment of a highly efficient virus-inducible CRISPR/Cas9 system in insect cells. *Antiviral Res.* **130**, 50–57.

Dong, Y., Simões, M. L., Marois, E. and Dimopoulos, G. (2018). CRISPR/Cas9 -mediated gene knockout of *Anopheles gambiae FREP1* suppresses malaria parasite infection. *PLoS Pathog.* **14**, e1006898.

Doudna, J. A. and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P. and Lieberman Aiden, E. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome- length scaffolds. *Science* **356**, 92–95.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Turner, S., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138.

Ellegren, H. (2011). Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. *Genetics* **12**, 157–166.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C. and Gibbs, R. A. (2012). Mind the Gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768.

Ewen-Campen, B. and Perrimon, N. (2018). *ovoD* co-selection: a method for enriching CRISPR/Cas9-edited alleles in *Drosophila*. *bioRxiv* 310854.

Fauci, A. S. and Morens, D. M. (2016). Zika Virus in the Americas — Yet Another Arbovirus Threat. *N. Engl. J. Med.* **374**, 601–604.

Faull, K. J. and Williams, C. R. (2015). Intraspecific variation in desiccation survival time of Aedes aegypti (L.) mosquito eggs of Australian origin. *J. Vector Ecol.* **40**, 292–300.

Flores, H. A. and O'Neill, S. L. (2018). Controlling vector-borne diseases by releasing modified mosquitoes. *Nat. Rev. Microbiol.* **16**, 508–518.

Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517.

Fontaine, A., Filipović, I., Fansiri, T., Hoffmann, A. A., Cheng, C., Kirkpatrick, M., Rašić, G. and Lambrechts, L. (2017). Extensive genetic differentiation

between homomorphic sex chromosomes in the mosquito vector, *Aedes aegypti*. *Genome Biol. Evol.* **9**, 2322–2335.

Fu, G., Condon, K. C., Epton, M. J., Gong, P., Jin, L., Condon, G. C., Morrison, N. I., Dafa'alla, T. H. and Alphey, L. (2007). Female-specific insect lethality engineered using alternative splicing. *Nat. Biotechnol.* **25**, 353–357.

Galizi, R., Doyle, L. A., Menichelli, M., Bernardini, F., Deredec, A., Burt, A., Stoddard, B. L., Windbichler, N. and Crisanti, A. (2014). A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nat. Commun.* **5**, 3977.

Galizi, R., Hammond, A., Kyrou, K., Taxiarchi, C., Bernardini, F., O'Loughlin, S. M., Papathanos, P. A., Nolan, T., Windbichler, N. and Crisanti, A. (2016). A CRISPR-Cas9 sex-ratio distortion system for genetic control. *Sci. Rep.* **6**, 31139.

Gammon, D. B. and Mello, C. C. (2015). RNA interference-mediated antiviral defense in insects. *Curr. Opin. Insect Sci.* **8**, 111–120.

Gantz, V. M. and Akbari, O. S. (2018). Gene editing technologies and applications for insects. *Curr. Opin. Insect Sci.* In press.

Gantz, V. M., Jasinskiene, N., Tatarenkova, O., Fazekas, A., Macias, V. M., Bier, E. and James, A. A. (2015). Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. *Proc. Natl. Acad. Sci.* **112**, E6736–E67435.

Gilchrist, B. M. and Haldane, J. B. S. (1947). Sex linkage and sex determination in a mosquito, *Culex molestus*. *Hereditas* **33**, 175–190.

Gilles, J. R. L., Schetelig, M. F., Scolari, F., Marec, F., Capurro, M. L., Franz, G. and Bourtzis, K. (2014). Towards mosquito sterile insect technique programmes: exploring genetic, molecular, mechanical and behavioural methods of sex separation in mosquitoes. *Acta Trop.* **132**, S178-187.

Gilles, A. F., Schinko, J. B. and Averof, M. (2015). Efficient CRISPR-mediated gene targeting and transgene replacement in the beetle *Tribolium castaneum*. *Development* **142**, 2832–2839.

Giraldo-Calderón, G. I., Emrich, S. J., MacCallum, R. M., Maslen, G., Emrich, S., Collins, F., Dialynas, E., Topalis, P., Ho, N., Gesing, S., Madey, G., Collins, F. H., Lawson, D., Kersey, P., Allen, J., Christensen, M., Hughes, D., Koscielny, G., Langridge, N., Gallego, E. L., Megy, K., Wilson, D., Gelbart, B., Emmert, D., Russo, S., Zhou, P., Christophides, G., Brockman, A., Kirmitzoglou, I., MacCallum, B., Tiirikka, T., Louis, K., Dritsou, V., Mitraka, E., Werner-Washburn, M., Baker, P., Platero, H., Aguilar, A., Bogol, S., Campbell, D., Carmichael, R., Cieslak, D., Davis, G., Konopinski, N., Nabrzyski, J., Reinking, C., Sheehan, A., Szakonyi, S. and Wieck, R. (2015). VectorBase: An updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* **43**, D707–D713.

Gloria-Soria, A., Ayala, D., Bheecarry, A., Calderon-Arguedas, O., Chadee, D. D., Chiappero, M., Coetzee, M., Elahee, K. Bin, Fernandez-Salas, I., Kamal, H. A., Kamgang, B., Khater, E. I. M., Kramer, L. D., Kramer, V., Lopez-Solis, A., Lutomiah, J., Martins, A., Micieli, M. V., Paupy, C., Ponlawat, A., Rahola, N., Rasheed, S. B., Richardson, J. B., Saleh, A. A., Sanchez-Casas, R. M., Seixas, G., Sousa, C. A., Tabachnick, W. J., Troyo, A. and Powell, J. R. (2016). Global genetic diversity of Aedes aegypti. *Mol. Ecol.* **25**, 5377–5395.

Gong, P., Epton, M. J., Fu, G., Scaife, S., Hiscox, A., Condon, K. C., Condon, G. C., Morrison, N. I., Kelly, D. W., Dafa'Alla, T., Coleman, P. G. and Alphey, L. (2005). A dominant lethal genetic system for autocidal control of the Mediterranean fruitfly. *Nat. Biotechnol.* **23**, 453–456.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A. (2011). Full-

length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

Gratz, S. J., Ukken, F. P., Rubinstein, C. D., Thiede, G., Donohue, L. K., Cummings, A. M. and O'Connor-Giles, K. M. (2014). Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in *Drosophila*. *Genetics* **196**, 961–971.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075.

Hall, A. B., Qi, Y., Timoshevskiy, V., Sharakhova, M. V., Sharakhov, I. V. and Tu, Z. (2013). Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics* **14**,.

Hall, A. B., Timoshevskiy, V. A., Sharakhova, M. V., Jiang, X., Basu, S., Anderson, M. A. E., Hu, W., Sharakhov, I. V., Adelman, Z. N. and Tu, Z. (2014). Insights into the preservation of the homomorphic sex-determining chromosome of *Aedes aegypti* from the discovery of a male-biased gene tightly linked to the M-locus. *Genome Biol. Evol.* **6**, 179–191.

Hall, A. B., Basu, S., Jiang, X., Qi, Y., Timoshevskiy, V. A., Biedler, J. K., Sharakhova, M. V, Elahi, R., Anderson, M. A. E., Chen, X., Sharakhov, I. V, Adelman, Z. N. and Tu, Z. (2015). A male-determining factor in the mosquito *Aedes aegypti*. *Science* **348**, 1268–1270.

Hall, A. B., Papathanos, P.-A., Sharma, A., Cheng, C., Akbari, O. S., Assour, L., Bergman, N. H., Cagnetti, A., Crisanti, A., Dottorini, T., Fiorentini, E., Galizi, R., Hnath, J., Jiang, X., Koren, S., Nolan, T., Radune, D., Sharakhova, M. V., Steele, A., Timoshevskiy, V. A., Windbichler, N., Zhang, S., Hahn, M. W., Phillippy, A. M., Emrich, S. J., Sharakhov, I. V., Tu, Z. J. and Besansky, N. J. (2016). Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. *Proc. Natl. Acad. Sci.* **113**, 201525164.

Hallmann, C. A., Foppen, R. P. B., Van Turnhout, C. A. M., De Kroon, H. and

Jongejans, E. (2014). Declines in insectivorous birds are associated with high neonicotinoid concentrations. *Nature* **511**, 341–343.

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., Stenmans, W., Müller, A., Sumser, H., Hörren, T., Goulson, D. and de Kroon, H. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS One* **12**, e0185809.

Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., Katsanos, D., Gribble, M., Baker, D., Marois, E., Russell, S., Burt, A., Windbichler, N., Crisanti, A. and Nolan, T. (2016). A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nat. Biotechnol.* **34**, 78.

Handler, A. M. (2002). Use of the *piggyBac* transposon for germ-line transformation of insects. *Insect Biochem. Mol. Biol.* **32**, 1211–1220.

Harkess, A., Zhou, J., Xu, C., Bowers, J. E., Van Der Hulst, R., Ayyampalayam, S., Mercati, F., Riccardi, P., McKain, M. R., Kakrana, A., Tang, H., Ray, J., Groenendijk, J., Arikit, S., Mathioni, S. M., Nakano, M., Shan, H., Telgmann-Rauber, A., Kanno, A., Yue, Z., Chen, H., Li, W., Chen, Y., Xu, X., Zhang, Y., Luo, S., Chen, H., Gao, J., Mao, Z., Pires, J. C., Luo, M., Kudrna, D., Wing, R. A., Meyers, B. C., Yi, K., Kong, H., Lavrijsen, P., Sunseri, F., Falavigna, A., Ye, Y., Leebens-Mack, J. H. and Chen, G. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat. Commun.* **8**, 1279.

Harmon, A. Tweets Inspire Rival Scientists to Come Together to Fight Zika. *New York Times* 1 Apr 2016: A1.

Harris, A. F., Nimmo, D., McKemey, A. R., Kelly, N., Scaife, S., Donnelly, C. A., Beech, C., Petrie, W. D. and Alphey, L. (2011). Field performance of engineered male mosquitoes. *Nat Biotech* **29**, 1034–1037.

Henking, H. (1891). Untersuchungen uber die ersten Entwicklungsvorigoenge in den Eiern der Insektern II Uber spermatogenese und Beziehung zur Eientwicklung bei *Pyrrhocoris apterus*. *Zeitschrift für wissenschaftliche Zool.* **51**, 685–736.

Hickey, W. A. and Craig, G. B. (1966). Genetic distortion of sex ratio in a mosquito, *Aedes aegypti*. *Genetics* **53**, 1177–1196.

Hiscox, A., Kaye, A., Vongphayloth, K., Banks, I., Piffer, M., Khammanithong, P., Sananikhom, P., Kaul, S., Hill, N., Lindsay, S. W. and Brey, P. T. (2013). Risk factors for the presence of *Aedes aegypti* and *Aedes albopictus* in domestic water-holding containers in areas impacted by the Nam Theun 2 hydroelectric project, Laos. *Am. J. Trop. Med. Hyg.* **88**, 1070–1078.

Hoang, K. P., Teo, T. M., Ho, T. X. and Le, V. S. (2016). Mechanisms of sex determination and transmission ratio distortion in *Aedes aegypti*. *Parasit. Vectors* **9**, 49.

Hoffmann, A., Montgomery, B. L., Popovici, J., Iturbe-Ormaetxe, I., Johnson, P. H., Muzzi, F., Greenfield, M., Durkan, M., Leong, Y. S., Dong, Y., Cook, H., Axford, J., Callahan, A. G., Kenny, N., Omodei, C., McGraw, E. A., Ryan, P. A., Ritchie, S. A., Turelli, M. and O'Neill, S. L. (2011). Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission. *Nature* **476**, 454–457.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, M. C., Wides, R., Salzberg, S. L., Loftus, B., Hoffman, S. L., et al. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149.

Hsu, P. D., Lander, E. S. and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278.

Huang, Y., Chen, Y., Zeng, B., Wang, Y., James, A. A., Gurr, G. M., Yang, G., Lin, X., Huang, Y. and You, M. (2016). CRISPR/Cas9 mediated knockout of the abdominal-A homeotic gene in the global pest, diamondback moth (*Plutella xylostella*). *Insect Biochem. Mol. Biol.* **75**, 98–106.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.

Itokawa, K., Komagata, O., Kasai, S., Ogawa, K. and Tomita, T. (2016). Testing the causality between *CYP9M10* and pyrethroid resistance using the TALEN

and CRISPR/Cas9 technologies. *Sci. Rep.* **6**, 1–10.

Jasinskiene, N., Coates, C. J., Benedict, M. Q., Cornel, A. J., Rafferty, C. S., James, A. A. and Collins, F. H. (1998). Stable transformation of the yellow fever mosquito, *Aedes aegypti*, with the *Hermes* element from the housefly. *Proc. Natl. Acad. Sci.* **95**, 3743–3747.

Jiggins, F. M. (2017). The spread of *Wolbachia* through mosquito populations. *PLoS Biol.* **15**, e2002780.

Jiménez, L. V., Kang, B. K., DeBruyn, B., Lovin, D. D. and Severson, D. W. (2004). Characterization of an *Aedes aegypti* bacterial artificial chromosome (BAC) library and chromosomal assignment of BAC clones for physical mapping quantitative trait loci that influence *Plasmodium* susceptibility. *Insect Mol. Biol.* **13**, 37–44.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–822.

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A. T., Charpentier, E., Nogales, E. and Doudna, J. A. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997.

Jordan, I. K. and McDonald, J. F. (1998). Evolution of the *copia* retrotransposon in the *Drosophila melanogaster* species subgroup. *Mol. Biol. Evol.* **15**, 1160–1171.

Kaiser, V. B. and Bachtrog, D. (2010). Evolution of sex chromosomes in insects. *Annu. Rev. Genet.* **44**, 91–112.

Kapun, M., Van Schalkwyk, H., McAllister, B., Flatt, T. and Schlötterer, C. (2014). Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of *Drosophila melanogaster*. *Mol. Ecol.* **23**, 1813–1827.

Kejnovsky, E., Hobza, R., Cermak, T., Kubat, Z. and Vyskot, B. (2009). The role of repetitive DNA in structure and evolution of sex chromosomes in plants.

Heredity (Edinb). **102**, 533–541.

Keller, O., Kollmar, M., Stanke, M. and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763.

Kent, W. J. (2002). BLAT — The BLAST -Like Alignment Tool. *Genome Res.* **12**, 656–664.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493.

Kim, A., Terzian, C., Santamaria, P., Pélisson, A., Purd'homme, N. and Bucheton, A. (1994). Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **91**, 1285–1289.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Kistler, K. E., Vosshall, L. B. and Matthews, B. J. (2015). Genome engineering with CRISPR-Cas9 in the mosquito *Aedes aegypti*. *Cell Rep.* **11**, 51–60.

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., Gonzales, A. P. W., Li, Z., Peterson, R. T., Yeh, J.-R. J., Aryee, M. J. and Joung, J. K. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485.

Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z. and Keith Joung, J. (2016). High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495.

Kohno, H., Suenami, S., Takeuchi, H., Sasaki, T. and Kubo, T. (2016). Production of knockout mutants by CRISPR/Cas9 in the European Honeybee, *Apis mellifera* L. *Zoolog. Sci.* **33**, 505–512.

Kondo, S. and Ueda, R. (2013). Highly improved gene targeting by germline-specific Cas9 expression in *Drosophila*. *Genetics* **195**, 715–721.

Koren, S. and Phillippy, A. M. (2015). One chromosome, one contig: complete

microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **23**, 110–120.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J., Ganapathy, G., Wang, Z., Rasko, D. A., Mccombie, W. R., Jarvis, D. and Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700.

Koren, S., Rhie, A., Walenz, B. P., Dilthey, A. T., Bickhart, D. M., Kingan, S. B., Hiendleder, S., Williams, J. L., Smith, T. P. L. and Phillippy, A. (2018). Complete assembly of parental haplotypes with trio binning. *bioRxiv* 271486.

Korlach, J., Gedman, G., Kingan, S. B., Chin, C. S., Howard, J. T., Audet, J. N., Cantin, L. and Jarvis, E. D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience* **6**, 1–16.

Kraemer, M. U. G., Sinka, M. E., Duda, K. A., Mylne, A., Shearer, F. M., Barker, C. M., Moore, C. G., Carvalho, R. G., Coelho, G. E., Van Bortel, W., Hendrickx, G., Schaffner, F., Elyazar, I. R., Teng, H.-J., Brady, O. J., Messina, J. P., Pigott, D. M., Scott, T. W., Smith, D. L., Wint, G. W., Golding, N. and Hay, S. I. (2015). The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. *Elife* **4**, e08347.

Kroeger, A., Lenhart, A., Ochoa, M., Villegas, E., Levy, M., Alexander, N. and McCall, P. J. (2006). Effective control of dengue vectors with curtains and water container covers treated with insecticide in Mexico and Venezuela: cluster randomised trials. *BMJ* **332**, 1247.

Krzywinska, E., Dennison, N. J., Lycett, G. J. and Krzywinski, J. (2016). A maleness gene in the malaria mosquito *Anopheles gambiae*. *Science* **353**, 67–69.

Kudoh, T., Takahashi, M., Osabe, T., Toyoda, A., Hirakawa, H., Suzuki, Y., Ohmido, N. and Onodera, Y. (2018). Molecular insights into the non-recombining nature of the spinach male-determining region. *Mol. Genet. Genomics* **293**, 557–568.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu,

C. and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12.

Kyle, J. L. and Harris, E. (2008). Global spread and persistence of dengue. *Annu. Rev. Microbiol.* **62**, 71–92.

Kyrou, K., Hammond, A. M., Galizi, R., Kranjc, N., Burt, A., Beaghton, A. K., Nolan, T. and Crisanti, A. (2018). A CRISPR–Cas9 gene drive targeting *doublesex* causes complete population suppression in caged Anopheles gambiae mosquitoes. *Nat. Biotechnol.* **36**, 1062–1066.

Labbé, G. M. C., Nimmo, D. D. and Alphey, L. (2010). *Piggybac*- and PhiC31-mediated genetic transformation of the Asian tiger mosquito, *Aedes albopictus* (Skuse). *PLoS Negl. Trop. Dis.* **4**, e788.

Lacroix, R., McKemey, A. R., Raduan, N., Kwee Wee, L., Hong Ming, W., Guat Ney, T., Siti Rahidah, A. A., Salman, S., Subramaniam, S., Nordin, O., Norhaida Hanum, A. T., Angamuthu, C., Marlina Mansor, S., Lees, R. S., Naish, N., Scaife, S., Gray, P., Labbé, G., Beech, C., Nimmo, D., Alphey, L., Vasan, S. S., Han Lim, L., Wasi A., N. and Murad, S. (2012). Open field release of genetically engineered sterile male *Aedes aegypti* in Malaysia. *PLoS One* **7**, e42771.

Lambert, B., North, A., Burt, A. and Godfray, H. C. J. (2018). The use of driving endonuclease genes to suppress mosquito vectors of malaria in temporally variable environments. *Malar. J.* **17**, 154.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

Lapinaite, A., Doudna, J. A. and Cate, J. H. D. (2018). Programmable RNA recognition using a CRISPR-associated Argonaute. *Proc. Natl. Acad. Sci.* **115**, 3368–3373.

Laughlin, C. A., Morens, D. M., Cassetti, M. C., Costero-Saint Denis, A., San

Martin, J. L., Whitehead, S. S. and Fauci, A. S. (2012). Dengue research opportunities in the Americas. *J. Infect. Dis.* **206**, 1121–1127.

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., van Sluys, M.-A., Soltis, P. S., Xu, X., Yang, H. and Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* **115**, 4325–4333.

Lewis, S. H., Quarles, K. A., Yang, Y., Tanguy, M., Frézal, L., Smith, S. A., Sharma, P. P., Cordaux, R., Gilbert, C., Giraud, I., Collins, D. H., Zamore, P. D., Miska, E. A., Sarkies, P. and Jiggins, F. M. (2018). Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. *Nat. Ecol. Evol.* **2**, 174–181.

Li, F. and Scott, M. J. (2016). CRISPR/Cas9-mediated mutagenesis of the *white* and *Sex lethal* loci in the invasive pest, Drosophila suzukii. *Biochem. Biophys. Res. Commun.* **469**, 911–916.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K. and Wang, J. (2009b). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967.

Li, Y., Zhang, J., Chen, D., Yang, P., Jiang, F., Wang, X. and Kang, L. (2016). CRISPR/Cas9 in locusts: Successful establishment of an olfactory deficiency line by targeting the mutagenesis of an odorant receptor co-receptor (Orco). *Insect Biochem. Mol. Biol.* **79**, 27–35.

Li, M., Au, L. Y. C., Douglah, D., Chong, A., White, B. J., Ferree, P. M. and Akbari, O. S. (2017a). Generation of heritable germline mutations in the jewel wasp *Nasonia vitripennis* using CRISPR/Cas9. *Sci. Rep.* **7**, 901.

Li, M., Bui, M., Yang, T., Bowman, C. S., White, B. J. and Akbari, O. S. (2017b). Germline Cas9 expression yields highly efficient genome engineering in a major worldwide disease vector, *Aedes aegypti*. *Proc. Natl. Acad. Sci.* **114**, E10540–E10549.

Li, M., Akbari, O. S. and White, B. J. (2018). Highly efficient site-specific mutagenesis in malaria mosquitoes using CRISPR. *G3* **8**, 653–658.

Lieberman-Aiden, E., Berkum, N. L. Van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J. and Mirny, L. A. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the Human Genome. *Science* **326**, 289–293.

Lin, C.-C. and Potter, C. J. (2016). Non-Mendelian dominant maternal effects caused by CRISPR/Cas9 transgenic components in *Drosophila melanogaster*. *G3* **6**, 3685–3691.

Liu, Z., Moore, P. H., Ma, H., Ackerman, C. M., Ragiba, M., Yu, Q., Pearl, H. M., Kim, M. S., Charlton, J. W., Stiles, J. I., Zee, F. T., Paterson, A. H. and Ming, R. (2004). A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature* **427**, 348.

Lobo, N. F., Clayton, J. R., Fraser, M. J., Kafatos, F. C. and Collins, F. H. (2006). High efficiency germ-line transformation of mosquitoes. *Nat. Protoc.* **1**, 1312–1317.

Macias, V. M., Ohm, J. R. and Rasgon, J. L. (2017). Gene drive for mosquito control: Where did it come from and where are we headed? *Int. J. Environ. Res. Public Health* **14**, 1006.

Mahajan, S. and Bachtrog, D. (2017). Convergent evolution of Y chromosome gene content in flies. *Nat. Commun.* **8**, 785.

Mahajan, S., Wei, K. H. C., Nalley, M. J., Gibilisco, L. and Bachtrog, D. (2018). De novo assembly of a young *Drosophila* Y chromosome using single-molecule sequencing and chromatin conformation capture. *PLoS Biol.* **16**, e2006348.

Malone, C. D. and Hannon, G. J. (2009). Small RNAs as Guardians of the Genome. *Cell* **136**, 656–668.

Marais, G. A. B., Nicolas, M., Bergero, R., Chambrier, P., Kejnovsky, E., Monéger, F., Hobza, R., Widmer, A. and Charlesworth, D. (2008). Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. *Curr. Biol.* **18**, 545–549.

Maselko, M., Heinsch, S., Das, S. and Smanski, M. J. (2018). Genetic incompatibility combined with female-lethality is effective and robust in simulations of *Aedes aegypti* population control. *bioRxiv* 316406.

Matthews, B. J., McBride, C. S., DeGennaro, M., Despo, O. and Vosshall, L. B. (2016). The neurotranscriptome of the *Aedes aegypti* mosquito. *BMC Genomics* **17**, 32.

Matthews, B. J., Dudchenko, O., Kingan, S. B., Koren, S., Antoshechkin, I., Crawford, J. E., Glassford, W. J., Herre, M., Redmond, S. N., Rose, N. H., Weedall, G. D., Wu, Y., Vosshall, L. B., et al. (2018). Improved reference genome of *Aedes aegypti* informs arbovirus vector control. *Nature* **563**, 501–507.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46.

Miesen, P., Ivens, A., Buck, A. H. and van Rij, R. P. (2016). Small RNA profiling in Dengue Virus 2-infected *Aedes* mosquito cells reveals viral piRNAs and novel host miRNAs. *PLoS Negl. Trop. Dis.* **10**, 1–22.

Miller, J. R., Koren, S., Dilley, K. A., Puri, V., Brown, D. M., Harkins, D. M., Thibaud-Nissen, F., Rosen, B., Chen, X.-G., Tu, Z., Sharakhov, I. V, Sharakhova, M. V, Sebra, R., Stockwell, T. B., Bergman, N. H., Sutton, G. G., Phillippy, A. M., Piermarini, P. M. and Shabman, R. S. (2018). Analysis of the *Aedes albopictus* C6/36 genome provides insight into cell line utility for viral propagation. *Gigascience* **7**, 1–13.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Flouri, T., Beutel, R. G., Niehuis, O. and Petersen, M. (2014). Phylogenomics resolves the timing and

pattern of insect evolution. *Science* **346**, 763–768.

Moreira, L. A., Iturbe-Ormaetxe, I., Jeffery, J. A., Lu, G., Pyke, A. T., Hedges, L. M., Rocha, B. C., Hall-Mendelin, S., Day, A., Riegler, M., Hugo, L. E., Johnson, K. N., Kay, B. H., McGraw, E. A., van den Hurk, A. F., Ryan, P. A. and O'Neill, S. L. (2009). A *Wolbachia* symbiont in *Aedes aegypti* limits infection with dengue, chikungunya, and *Plasmodium*. *Cell* **139**, 1268–1278.

Motara, M. A. and Rai, K. S. (1977). Chromosomal differentiation in two species of *Aedes* and their hybrids revealed by giemsa C-banding. *Chromosoma* **64**, 125–132.

Moyes, C. L., Vontas, J., Martins, A. J., Ng, L. C., Koou, S. Y., Dusfour, I., Raghavendra, K., Pinto, J., Corbel, V., David, J. P. and Weetman, D. (2017). Contemporary status of insecticide resistance in the major *Aedes* vectors of arboviruses infecting humans. *PLoS Negl. Trop. Dis.* **11**, 1–20.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutat. Res.* **1**, 2–9.

Musso, D., Cao-Lormeau, V. M. and Gubler, D. J. (2015). Zika virus: following the path of dengue and chikungunya? *Lancet* **386**, 243–244.

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Besansky, N. J., et al. (2015). Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. *Science* **347**, 1258522.

Nene, V., Wortman, J. R., Lawson, D., Haas, B., Kodira, C., Tu, Z. J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., Ren, Q., Zdobnov, E. M., Severson, D. W., et al. (2007). Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316**, 1718–1723.

Newton, M. E., Southern, D. I. and Wood, R. J. (1974). X and Y chromosomes of *Aedes aegypti* (L.) distinguished by Giemsa C-banding. *Chromosoma* **49**, 41–49.

Newton, M. E., Wood, R. J. and Southern, D. I. (1978). Cytological mapping of the M and D loci in the mosquito, *Aedes aegypti* (L.). *Genetica* **48**, 137–143.

Nimmo, D. D., Alphey, L., Meredith, J. M. and Eggleston, P. (2006). High efficiency site-specific genetic engineering of the mosquito genome. *Insect Mol. Biol.* **15**, 129–136.

Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O. O., Gootenberg, J. S., Mori, H., Oura, S., Holmes, B., Tanaka, M., Seki, M., Hirano, H., Aburatani, H., Ishitani, R., Ikawa, M., Yachie, N., Zhang, F. and Nureki, O. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262.

Niu, B., Fu, L., Sun, S. and Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11**, 187.

O'Brochta, D. A. and Handler, A. M. (2008). Perspectives on the state of insect transgenics. In *Transgenesis and the Management of Vector-Borne Disease. Advances in Experimental Medicine and Biology* (ed. Aksoy, S.), pp. 1–18. New York: Springer.

O'Connell, M. R., Oakes, B. L., Sternberg, S. H., East-Seletsky, A., Kaplan, M. and Doudna, J. A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature* **516**, 263–266.

Otto, S. P. (2009). The evolutionary enigma of sex. *Am. Nat.* **174**, S1–S14.

Overcash, J. M., Aryan, A., Myles, K. M. and Adelman, Z. N. (2015). Understanding the DNA damage response in order to achieve desired gene editing outcomes in mosquitoes. *Chromosom. Res.* **23**, 31–42.

Oye, K. A., Esvelt, K., Appleton, E., Catteruccia, F., Church, G., Kuiken, T., Lightfoot, S. B.-Y., McNamara, J., Smidler, A. and Collins, J. P. (2014). Regulating gene drives. *Science* **345**, 626–628.

Palatini, U., Miesen, P., Carballar-Lejarazu, R., Ometto, L., Rizzo, E., Tu, Z., van Rij, R. P. and Bonizzoni, M. (2017). Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors *Aedes aegypti* and *Aedes albopictus*. *BMC Genomics* **18**, 1–15.

Pang, T., Mak, T. K. and Gubler, D. J. (2017). Prevention and control of dengue—the light at the end of the tunnel. *Lancet Infect. Dis.* **17**, e79–e87.

Papathanos, P. A., Bossin, H. C., Benedict, M. Q., Catteruccia, F., Malcolm, C. A., Alphey, L. and Crisanti, A. (2009). Sex separation strategies: Past experience and new approaches. *Malar. J.* **8**, S5.

Paredes-Esquivel, C., Lenhart, A., del Río, R., Leza, M. M., Estrugo, M., Chalco, E., Casanova, W. and Miranda, M. Á. (2016). The impact of indoor residual spraying of deltamethrin on dengue vector populations in the Peruvian Amazon. *Acta Trop.* **154**, 139–144.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J. and Quackenbush, J. (2003). TIGR gene indices clustering tools (TGICL): A software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651–652.

Phuc, H. K., Andreasen, M. H., Burton, R. S., Vass, C., Epton, M. J., Pape, G., Fu, G., Condon, K. C., Scaife, S., Donnelly, C. A., Coleman, P. G., White-Cooper, H. and Alphey, L. (2007). Late-acting dominant lethal genetic systems and mosquito control. *BMC Biol.* **5**, 11.

Port, F., Chen, H.-M., Lee, T. and Bullock, S. L. (2014). Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in *Drosophila*. *Proc. Natl. Acad. Sci.* **111**, E2967–E2076.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680.

Reidenbach, K. R., Cook, S., Bertone, M. A., Harbach, R. E., Wiegmann, B. M. and Besansky, N. J. (2009). Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: Culicidae) based on nuclear genes and morphology. *BMC Evol. Biol.* **9**, 298.

Ren, X., Sun, J., Housden, B. E., Hu, Y., Roesel, C., Lin, S., Liu, L.-P., Yang, Z., Mao, D., Sun, L., Wu, Q., Ji, J.-Y., Xi, J., Mohr, S. E., Xu, J.,

Perrimon, N. and Ni, J.-Q. (2013). Optimized gene editing technology for *Drosophila melanogaster* using germ line-specific Cas9. *Proc. Natl. Acad. Sci.* **110**, 19012–19017.

Ren, X., Yang, Z., Xu, J., Sun, J., Mao, D., Hu, Y., Yang, S.-J., Qiao, H.-H., Wang, X., Hu, Q., Deng, P., Liu, L.-P., Ji, J.-Y., Li, J. B. and Ni, J.-Q. (2014). Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in *Drosophila*. *Cell Rep.* **9**, 1151–1162.

Rice, W. R. (1987). Genetic Hitchhiking and the Evolution of Reduced Genetic Activity of the Y Sex Chromosome. *Genetics* **116**, 161–167.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. and Mesirov, J. P. (2011a). Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26.

Robinson, G. E., Hackett, K. J., Purcell-Miramontes, M., Brown, S. J., Evans, J. D., Goldsmith, M. R., Lawson, D., Okamuro, J., Robertson, H. M. and Schneider, D. J. (2011b). Creating a Buzz About Insect Genomes. *Science* **331**, 1386.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C. and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51.

Salvemini, M., Mauro, U., Lombardo, F., Milano, A., Zazzaro, V., Arcà, B., Polito, L. C. and Saccone, G. (2011). Genomic organization and splicing evolution of the *doublesex* gene, a *Drosophila* regulator of sexual differentiation, in the dengue and yellow fever mosquito *Aedes aegypti*. *BMC Evol. Biol.* **11**, 41.

Salvemini, M., D'Amato, R., Petrella, V., Aceto, S., Nimmo, D., Neira, M., Alphey, L., Polito, L. C. and Saccone, G. (2013). The orthologue of the fruitfly sex behaviour gene *fruitless* in the mosquito *Aedes aegypti*: evolution of genomic organisation and alternative splicing. *PLoS One* **8**, e48554.

Schetelig, M. F. and Wimmer, E. A. (2011). Insect Transgenesis and the Sterile Insect Technique. In *Insect Biotechnology* (ed. Vilcinskas, A.), pp. 169–194.

Dordrecht: Springer Netherlands.

Schlötterer, C., Tobler, R., Kofler, R. and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**, 749–763.

Severson, D. W. and Behura, S. K. (2012). Mosquito genomics: progress and challenges. *Annu. Rev. Entomol.* **57**, 143–166.

Severson, D. W., DeBruyn, B., Lovin, D. D., Brown, S. E., Knudson, D. L. and Morlais, I. (2004). Comparative genome analysis of the yellow fever mosquito *Aedes aegypti* with *Drosophila melanogaster* and the malaria vector mosquito *Anopheles gambiae*. *J. Hered.* **95**, 103–113.

Sharakhova, M. V., Timoshevskiy, V. A., Yang, F., Demin, S. I., Severson, D. W. and Sharakhov, I. V. (2011). Imaginal discs - A new source of chromosomes for genome mapping of the yellow fever mosquito *Aedes aegypti*. *PLoS Negl. Trop. Dis.* **5**, e1335.

Sharma, A., Heinze, S. D., Wu, Y., Kohlbrenner, T., Morilla, I., Brunner, C., Wimmer, E. A., van de Zande, L., Robinson, M. D., Beukeboom, L. W. and Bopp, D. (2017). Male sex in houseflies is determined by *Mdmd*, a paralog of the generic splice factor gene *CWC22*. *Science* **356**, 642–645.

Shaw, W. R. and Catteruccia, F. (2018). Vector biology meets disease control: using basic research to fight vector-borne diseases. *Nat. Microbiol.*

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212.

Simmons, C. P., Farrar, J. J., Chau, N. V. V. and Wills, B. (2012). Dengue. *N. Engl. J. Med.* **366**, 1423–1432.

Sinkins, S. P. and Gould, F. (2006). Gene drive systems for insect disease vectors. *Nat. Rev. Genet.* **7**, 427–435.

Smit, A. F. A., Hubley, R. and Green, P. RepeatMasker. https://www.repeatmasker.org/.

Smith, R. C., Walter, M. F., Hice, R. H., O'Brochta, D. A. and Atkinson, P. W.

(2007). Testis-specific expression of the β2 tubulin promoter of *Aedes aegypti* and its application as a genetic sex-separation marker. *Insect Mol. Biol.* **16**, 61–71.

Stokstad, E. (2018). Researchers launch plan to sequence 66,000 species in the United Kingdom. But that's just a start. *Science* 1 Nov 2018.

Sun, X., Le, H. D., Wahlstrom, J. M. and Karpen, G. H. (2003). Sequence analysis of a functional *Drosophila* centromere. *Genome Res.* **13**, 182–194.

Sun, D., Guo, Z., Liu, Y. and Zhang, Y. (2017). Progress and prospects of CRISPR/Cas systems in insects and other arthropods. *Front. Physiol.* **8**, 608.

Sutton, E. R., Yu, Y., Shimeld, S. M., White-Cooper, H. and Alphey, L. (2016). Identification of genes for engineering the male germline of *Aedes aegypti* and *Ceratitis capitata*. *BMC Genomics* **17**, 948.

Taning, C. N. T., Van Eynde, B., Yu, N., Ma, S. and Smagghe, G. (2017). CRISPR/Cas9 in insects: Applications, best practices and biosafety concerns. *J. Insect Physiol.* **98**, 245–257.

Tao, Q., Wang, A. and Zhang, H. B. (2002). One large-insert plant-transformation-competent BIBAC library and three BAC libraries of Japonica rice for genome research in rice and other grasses. *Theor. Appl. Genet.* **105**, 1058–1066.

Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M. and Kitts, P. (2013). Eukaryotic Genome Annotation Pipeline. In *The NCBI Handbook [Internet]. 2nd edition.*, pp. 133–156. Bethesda, MD: National Center for Biotechnology Information (US).

Thomas, D. D., Donnelly, C. A., Wood, R. J. and Alphey, L. S. (2000). Insect population control using a dominant, repressible, lethal genetic system. *Science* **287**, 2474–2476.

Timoshevskiy, V. A., Severson, D. W., DeBruyn, B. S., Black, W. C., Sharakhov, I. V. and Sharakhova, M. V. (2013). An integrated linkage, chromosome, and genome map for the yellow fever mosquito *Aedes aegypti*. *PLoS Negl. Trop. Dis.* **7**, e2052.

Timoshevskiy, V. A., Kinney, N. A., DeBruyn, B. S., Mao, C., Tu, Z., Severson, D. W., Sharakhov, I. V and Sharakhova, M. V (2014). Genomic composition and evolution of *Aedes aegypti* chromosomes revealed by the analysis of physically mapped supercontigs. *BMC Biol.* **12**, 27.

Tomchaney, M., Mysore, K., Sun, L., Li, P., Emrich, S. J., Severson, D. W. and Duman-Scheel, M. (2014). Examination of the genetic basis for sexual dimorphism in the *Aedes aegypti* (dengue vector mosquito) pupal brain. *Biol. Sex Differ.* **5**, 10.

Totten, D. C., Vuong, M., Litvinova, O. V, Jinwal, U. K., Gulia-Nuss, M., Harrell, R. A. and Beneš, H. (2013). Targeting gene expression to the female larval fat body of transgenic *Aedes aegypti* mosquitoes. *Insect Mol. Biol.* **22**, 18–30.

Toups, M. A. and Hahn, M. W. (2010). Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. *Genetics* **186**, 763–766.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53.

Tu, Z. and Coates, C. (2004). Mosquito transposable elements. *Insect Biochem. Mol. Biol.* **34**, 631–644.

Turner, J., Krishna, R., Van 't Hof, A. E., Sutton, E. R., Matzen, K. and Darby, A. C. (2018). The sequence of a male-specific genome region containing the sex determination switch in *Aedes aegypti*. *Parasit. Vectors* **11**, 549.

Verhulst, E. C. and van de Zande, L. (2015). Double nexus–*Doublesex* is the connecting element in sex determination. *Brief. Funct. Genomics* 1–11.

Vicoso, B. and Bachtrog, D. (2013). Reversal of an ancient sex chromosome to an autosome in *Drosophila*. *Nature* **499**, 332–335.

Vicoso, B. and Bachtrog, D. (2015). Numerous transitions of sex chromosomes in Diptera. *PLOS Biol.* **13**, e1002078.

Vicoso, B., Kaiser, V. B. and Bachtrog, D. (2013). Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proc. Natl. Acad. Sci.* **110**, 6453–6458.

Vinauger, C., Lahondère, C., Wolff, G. H., Locke, L. T., Liaw, J. E., Parrish, J. Z., Akbari, O. S., Dickinson, M. H. and Riffell, J. A. (2018). Modulation of host learning in *Aedes aegypti* mosquitoes. *Curr. Biol.* **28**, 333–344.

Vontas, J., Kioulos, E., Pavlidi, N., Morou, E., della Torre, A. and Ranson, H. (2012). Insecticide resistance in the major dengue vectors *Aedes albopictus* and *Aedes aegypti*. *Pestic. Biochem. Physiol.* **104**, 126–131.

Walker, T., Johnson, P. H., Moreira, L. A., Iturbe-Ormaetxe, I., Frentiu, F. D., McMeniman, C. J., Leong, Y. S., Dong, Y., Axford, J., Kriesner, P., Lloyd, A. L., Ritchie, S. A., O'Neill, S. L. and Hoffmann, A. A. (2011). The wMel *Wolbachia* strain blocks dengue and invades caged *Aedes aegypti* populations. *Nature* **476**, 450–455.

Wang, Y., Li, Z., Xu, J., Zeng, B., Ling, L., You, L., Chen, Y., Huang, Y. and Tan, A. (2013). The CRISPR/Cas System mediates efficient genome engineering in *Bombyx mori*. *Cell Res.* **23**, 1414–1416.

Wee, Y., Bhyan, S. B., Liu, Y., Lu, J., Li, X. and Zhao, M. (2018). The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Brief. Funct. Genomics* **00**, 1–12.

Weinert, L. A., Araujo-Jnr, E. V., Ahmed, M. Z. and Welch, J. J. (2015). The incidence of bacterial endosymbionts in terrestrial arthropods. *Proc. R. Soc. B Biol. Sci.* **282**, 3–8.

Whitfield, Z. J., Dolan, P. T., Kunitomi, M., Tassetto, M., Seetin, M. G., Oh, S., Heiner, C., Paxinos, E. and Andino, R. (2017). The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti* genome. *Curr. Biol.* **27**, 3511–3519.e7.

Wiegmann, B. M. and Richards, S. (2018). Genomes of Diptera. *Curr. Opin. Insect*

Sci. **25**, 116–124.

Windbichler, N., Menichelli, M., Papathanos, P. A., Thyme, S. B., Li, H., Ulge, U. Y., Hovde, B. T., Baker, D., Monnat, R. J., Burt, A. and Crisanti, A. (2011). A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. *Nature* **473**, 212–215.

Wise de Valdez, M. R., Suchman, E. L., Carlson, J. O. and Black, W. C. (2010). A large scale laboratory cage trial of *Aedes* Densonucleosis Virus (AeDNV). *J. Med. Entomol.* **47**, 392–399.

Wise de Valdez, M. R., Nimmo, D., Betz, J., Gong, H.-F., James, A. A., Alphey, L. and Black, W. C. (2011). Genetic elimination of dengue vector mosquitoes. *Proc. Natl. Acad. Sci.* **108**, 4772–4775.

Wong, L. H. and Choo, K. H. A. (2004). Evolutionary dynamics of transposable elements at the centromere. *Trends Genet.* **20**, 611–616.

World Health Organization (2009). *Dengue: Guidelines for diagnosis, treatment, prevention and control*.

World Health Organization (2016). Mosquito (vector) control emergency response and preparedness for Zika virus. http://www.who.int/neglected_diseases/news/mosqui.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134.

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., Mudivarti, P. A., Wyatt, P. W., Ji, H. P., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311.

Appendices

Appendix 1 Publications

Turner *et al. Parasites & Vectors* (2018) 11:549
https://doi.org/10.1186/s13071-018-3090-3

Parasites & Vectors

SHORT REPORT Open Access

The sequence of a male-specific genome region containing the sex determination switch in *Aedes aegypti*

Joe Turner[1,2†], Ritesh Krishna[1,3†], Arjen E. van't Hof[1,4†], Elizabeth R. Sutton[2,5,6], Kelly Matzen[2] and Alistair C. Darby[1*]

Abstract

Background: *Aedes aegypti* is the principal vector of several important arboviruses. Among the methods of vector control to limit transmission of disease are genetic strategies that involve the release of sterile or genetically modified non-biting males, which has generated interest in manipulating mosquito sex ratios. Sex determination in *Ae. aegypti* is controlled by a non-recombining Y chromosome-like region called the M locus, yet characterisation of this locus has been thwarted by the repetitive nature of the genome. In 2015, an M locus gene named *Nix* was identified that displays the qualities of a sex determination switch.

Results: With the use of a whole-genome bacterial artificial chromosome (BAC) library, we amplified and sequenced a ~200 kb region containing the male-determining gene *Nix*. In this study, we show that *Nix* is comprised of two exons separated by a 99 kb intron primarily composed of repetitive DNA, especially transposable elements.

Conclusions: *Nix*, an unusually large and highly repetitive gene, exhibits features in common with Y chromosome genes in other organisms. We speculate that the lack of recombination at the M locus has allowed the expansion of repeats in a manner characteristic of a sex-limited chromosome, in accordance with proposed models of sex chromosome evolution in insects.

Keywords: M locus, *Nix*, Sex determination, Chromosome evolution, Genomics, BAC, PacBio

Background

At least 2.5 billion people live in areas where they are at risk of dengue transmission from mosquitoes, principally *Ae. aegypti*, with an estimated 390 million infections per year [1, 2]. Recently, the emergence of chikungunya and Zika viruses further highlights the public health importance of *Ae. aegypti* [3, 4]. Future mosquito control strategies may incorporate genetic techniques such as the sustained release of sterile or transgenic "self-limiting" mosquitoes [5, 6]. Given that only female mosquitoes bite and spread disease, there has been substantial interest in manipulating mosquito sex determination using these genetic techniques and others, including gene drive [7, 8]. Therefore, elucidating the genetic basis for sex determination could, for instance, facilitate production of male-only cohorts for release, or allow transformation of mosquitoes with sex-specific "self-limiting" gene cassettes.

Sex determination in insects is variable, and generally not well understood outside of model species [9]. Unlike the malaria mosquito *Anopheles gambiae* and *Drosophila* species, *Ae. aegypti* does not have heteromorphic (XY) sex chromosomes [10]. Instead, the male phenotype is determined by a non-recombining M locus on one copy of autosome 1 [11–13]. This locus is poorly characterised because its highly repetitive nature has confounded attempts to study it based on the existing genome assembly [14]. The initial 1376 Mb *Ae. aegypti* reference genome was assembled from Sanger sequencing reads in 2007 [15], which are commonly not long enough to span the repetitive transposable elements that comprise a large proportion of the genome [16], and consequently the

* Correspondence: acdarby@liverpool.ac.uk
†Joe Turner, Ritesh Krishna and Arjen E. van't Hof contributed equally to this work.
[1]Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK
Full list of author information is available at the end of the article

assembly was relatively low quality [17]. Furthermore, the fact that both male and female genomic DNA was used for genome sequencing reduced the expected coverage of the M locus to one quarter of the autosome 1 sequences, further obscuring candidate M locus sequences [18].

Recently, a team of researchers was nevertheless able to identify *Nix*, a gene with male-specific, early embryonic expression. Knockout of *Nix* using CRISPR/Cas9 results in morphological feminisation of male mosquitoes along with feminisation of gene expression and female splice forms of the conserved sex-regulating genes *doublesex* (*dsx*) and *fruitless* (*fru*), strongly indicating that *Nix* is the upstream regulator of sexual differentiation [14]. The translated *Nix* protein contains two RNA recognition motifs and is hypothesised to be a splicing factor, acting either directly on *dsx* and *fru* or on currently unknown intermediates [19]. A comparison of sexually dimorphic gene expression in different mosquito tissue types also detected male-specific transcripts of *Nix* [20]. An ortholog of *Nix* is present in *Ae. albopictus*, but it is not known if the two are functionally homologous [21].

To date, *Nix* has only been characterised as an mRNA transcript. To fully understand this gene's role in sex determination and to utilise this knowledge for vector control, it is essential to decipher its genomic context. For this purpose, this study identifies and describes the region of the M locus in which *Nix* is located.

Results

Four BAC clones positive for *Nix* assembled into a single region of 207 kb with no gaps and a GC content of 40.2% (submitted to the NCBI as accession KY849907). The presence of the *Nix* gene in the assembled BACs was confirmed by BLASTN. The whole gene was present in tiled BACs, though not completely within individual BAC clones. Neither *Nix* nor the complete region could be found in the AaegL3 or Aag2 reference genome assemblies. The newly released AaegL5 male assembly contains *Nix* [22], and the assembled BACs aligned to the corresponding region in AaegL5 with > 99.9% identity, spanning a 2899 bp gap in the AaegL5 genome that is comprised mainly of repeats (Additional file 1: Figures S1, S2). While *Nix* was originally identified in the genome-sequenced Liverpool strain [14], PCR revealed that it is exclusively present in male genomic DNA from other geographically varied *Ae. aegypti* populations (Additional file 1: Figure S3), further strengthening the evidence that it is wholly present in the M locus.

The *Nix* gene was found to be made up of two exons with a single intron of 99 kb (Fig. 1). Although large introns are not uncommon in *Ae. aegypti* (average intron length ~5000 bp) [15], this intron is at the extreme end of intron sizes observed (Additional file 1: Figure S4), especially considering the small size of its protein coding regions (< 1000 bp). The gene structure is confirmed by Illumina RNA-Seq data clearly showing reads spanning

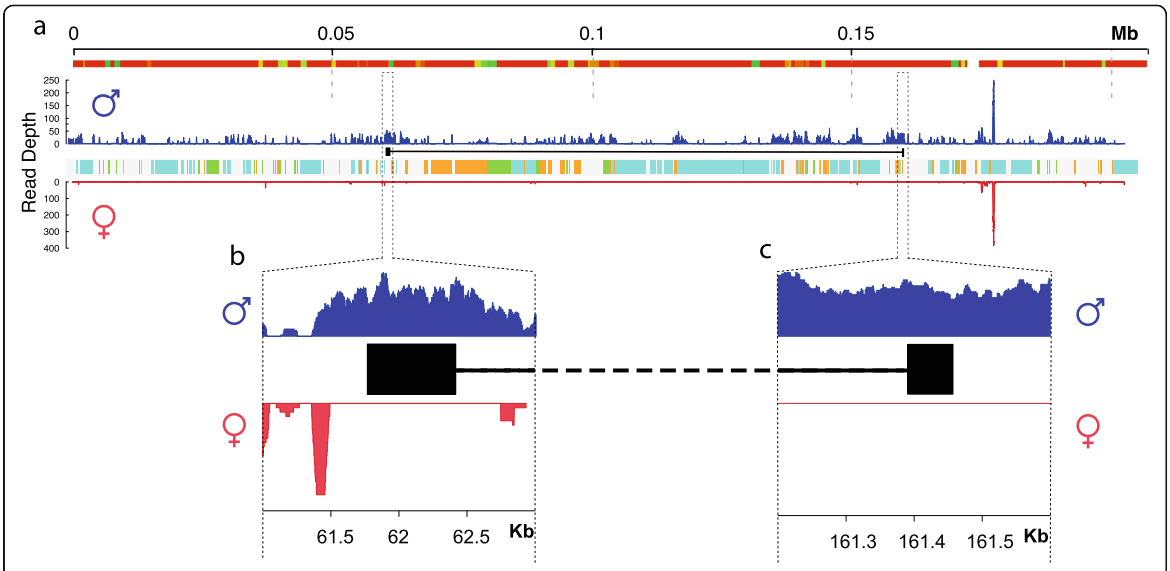


Fig. 1 Structure and gene expression of the ~207 kb genomic region containing the *Nix* gene. *Nix* is shown as two black boxes representing the exons, joined by a black line representing the intron. The top track of **a** shows the alignment of the sequence to the corresponding region of the reverse complement of the AaegL5 reference genome assembly, with colours representing percentage similarity (red: 100%; orange: > 90%; green: > 80%). Colours on the central track of **a** represent the classes of repetitive elements (orange: DNA transposons; cyan: Gypsy LTRs; green: Ty1/Copia LTRs). Blue histograms represent the coverage of RNA-Seq reads from male samples on the *y* axis; red histograms represent the coverage from female samples. **b** and **c** show enlargements of the first and second exons of *Nix* in the dotted regions in **a**, respectively

the intron between the two exons (Fig. 1). RepeatMasker identified approximately 55% of the sequenced region as repetitive, and the intron region of *Nix* as 72% repetitive (Additional file 2: Table S1).

Discussion

The genomic data from our assembled M locus region show that *Nix* is approximately 100 kb in length - exceptionally long even for an insect, and one of the longest in the mosquito genome. This is particularly unusual because *Nix* is expressed in early embryonic development, before the onset of the syncytial blastoderm stage 3–4 hours after oviposition [14], during which time most active genes have very short introns, or lack them entirely. There is evidence of selection against intron presence in genes expressed in the early *Ae. aegypti* zygote [23]. In *Drosophila*, the majority of early-expressed genes have small introns and encode small proteins, suggesting that selection has favoured high transcript turnover during early embryonic development due to the requirement for short cell cycles and rapid division [24]. It might therefore be expected that selection would limit the *Nix* intron's expansion to preserve efficient transcription in the zygote.

One possible explanation is the expansion of repetitive DNA. The RepeatMasker results reveal that the *Nix* region contains a high number of repetitive sequences, especially retrotransposons (Fig. 1, Additional file 2: Table S1). The M locus has accumulated repeats in between protein-coding DNA in a manner characteristic of a sex chromosome, which are prone to degeneration by Muller's ratchet due to the lack of recombination [25–27]. For instance, repetitive sequences comprise almost the entire *Anopheles gambiae* Y chromosome, and these repetitive sequences show rapid evolutionary divergence [28]. Similarly, certain Y chromosome genes of the plant *Silene latifolia* have much larger introns than their X chromosome copies due to the insertion of retrotransposons [29]. A more extreme version of this phenomenon is seen in *Drosophila,* where some Y chromosome genes, such as those involved in spermatogenesis, have gigantic repetitive introns, sometimes in the megabase range, that consequently make them many times larger than typical autosomal genes [30, 31].

It is therefore possible that the lack of recombination may pose constraints on the structure of the M locus, and in the absence of strong selection the *Nix* gene has degenerated outside the coding regions. Non-recombining sex loci such as the *Ae. aegypti* M locus may represent an evolutionary precursor to differentiated sex chromosomes, which are thought to emerge when sexually antagonistic alleles accumulate on either chromosome and favour reduced recombination between the two homologs, eventually leading to degeneration and loss of genes on the proto-Y [32]. Recent data appears to show that recombination is reduced along chromosome 1 even outside of the M locus [33], while the fully differentiated *Anopheles* X and Y chromosomes still display some degree of recombination with each other [28]. Thus, *Ae. aegypti* may be "further along" this evolutionary trajectory than previously assumed. The presence of additional repeats in our BAC assembly, which was obtained from the My1 mosquito strain, compared to the corresponding region in the AaegL5 genome assembly obtained from the Liverpool strain, suggests that the M locus may vary between strains outside of the *Nix* exons. Future work could investigate the population-level variation in the size and content of the M locus.

The *Ae. aegypti* M locus provides an intriguing example of the complexity of evolutionary forces acting on sex chromosomes, and further study of the locus will contribute to understanding the evolution of sex determination in insects and address general questions about the factors impacting gene and genome length. Importantly, these may also yield insights that can be applied to increase the efficiency of genetic strategies for vector control.

Methods

BAC library construction

A BAC library was constructed using living DH10b phage resistant *Escherichia coli* transfected with the pCC1BAC low copy number vector and *Ae. aegypti* genomic DNA from a DNA pool of approximately 50 sibling males (Amplicon Express, USA). Average insert size was 130 kb and the estimated coverage was ~5× for autosomal regions (~2.5× for sex specific regions). The male siblings were from one family from the My1 laboratory strain originating in Jinjang, Kuala Lumpur, Malaysia in the 1960s (described in [34]), after five generations of full-sib mating. Superpools and matrixpools were supplied to allow PCR based screening of the BAC library.

BAC library screening, isolation and sequencing

The BAC library was PCR screened using primers (Nix1F 3'-TTG AGT CTG AAA AGT CTA TGC AA-5', Nix1R 3'-TCG CTC TTC CGT GGC ATT TGA-5', Nix2F 3'-ACG TAG TCG GCA ACT CGA AG-5', Nix2R 3'-CTG GGA CAA ATC GAA CGG AA-5') based on the complete coding sequence of *Nix* (GenBank: KF732822). The first primer set was also used to screen for *Nix* in the genomic DNA of six male and six female individuals each from two wildtype *Ae. aegypti* strains.

Screening of the library resulted in four positive clones - two for each primer pair. These BAC clones were propagated, extracted using a Maxiprep kit (Qiagen, Hilden, Germany), pooled before SMRTbell library preparation (PacBio, Menlo Park, CA, USA), and sequenced on a single

SMRTcell using P6-C3 chemistry on the PacBio RS II platform (PacBio, USA).

Data analysis

The sequence data was trimmed to remove vector sequences and adaptors prior to assembly with the CANU v1 assembler [35], followed by sequence polishing with QUIVER.

BLASTN was used to assess the uniqueness of the assembled *Nix* region compared to the *Aedes aegypti* Liverpool reference genome AaegL3 and the newer Aag2 cell line assembly. Illumina data generated from male and female genomic DNA (accession numbers SRX290472 and SRX290470) and RNA (accession numbers SRX709698-SRX709703) were mapped to a combined reference containing the assembled *Nix* region added to the AaegL3 genome. DNA samples were mapped with BOWTIE 2.2.1 (using default parameters with -I 200 and -X 500) and RNA-Seq data with TOPHAT 2.1.1 version (using default parameters). RNA-Seq data was processed using the CUFFLINKS 2.2.1 pipeline to look for potential genes and male/female specific expression from the region.

Genes were predicted using AUGUSTUS and the *Aedes aegypti* model [15], repetitive regions described using REPEATMASKER 4.0.6 and the *Ae. aegypti* repeat database.

Additional files

Additional file 1: Figure S1. Alignment of the 207 kb BAC region to the corresponding region in the AaegL5 male reference assembly. **Figure S2.** Alignment of the 207 kb BAC region to chromosome 1 of the AaegL5 male reference assembly. **Figure S3.** PCR screening of the M locus gene *Nix* in male and female DNA of wild type *Aedes aegypti* strains. **Figure S4.** Intron size distribution in *Aedes aegypti* Liverpool reference genome AaegL3. (PDF 249 kb)

Additional file 2: Table S1. Types and abundance of repeats in the 207kb assembled M locus region and 99 kb *Nix* intron, identified by RepeatMasker using the *Aedes aegypti* repeat library. (XLSX 10 kb)

Abbreviations

AaegL#: *Aedes aegypti* Liverpool (LVP) strain reference genome assembly, version #; BAC: Bacterial artificial chromosome; CRISPR/Cas9: Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated protein-9 nuclease; LTR: Long terminal repeat; PCR: Polymerase chain reaction; RNA-Seq: RNA sequencing; WHO: World Health Organisation

Acknowledgments

PacBio sequencing was conducted at the Centre for Genomics Research, University of Liverpool with the assistance of Dr Margaret Hughes and Dr John Kenny. We thank Dr Andrea Betancourt and Dr Ilik Saccheri for comments on the manuscript.

Funding

This work was funded by UK Biotechnology and Biological Sciences Research Council (BBSRC) PhD training grant BB/M503460/1 (JT & ACD) and BBSRC grant BB/M001512/1 (KM & ACD).

Availability of data and materials

The assembly is available in NCBI GenBank under accession number KY849907 (https://www.ncbi.nlm.nih.gov/nuccore/KY849907). The FASTQ files for the RNA-Seq and genomic DNA reads used to map to the assembly are archived in the NCBI Sequence Read Archive (SRA) under the accession numbers SRX290472 and SRX290470 (genomic DNA) and SRX709698-SRX709703 (RNA).

Authors' contributions

JT, RK and AEvH contributed equally to this work. KM and ACD designed the study and obtained funding, with contribution from JT. KM provided mosquito samples. ERS and ACD commissioned the BAC library construction. AEvH and JT screened the BAC library and extracted DNA. AEvH performed BAC scaffolding. ACD oversaw sequencing and assembled the DNA sequence. RK performed the mapping and developed computational strategies for data analysis. JT performed the repeat masking. JT and ACD wrote the paper, with contribution from AEvH. JT, RK and ACD produced the figures. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

JT is a sponsored student (through the BBSRC Industrial CASE studentship) and KM is an employee of Oxitec Ltd., respectively, which therefore provided stipend or salary and other support for the research program.

Publisher's Note

Author details

[1]Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK. [2]Oxitec Ltd., 71 Innovation Drive, Milton Park, Abingdon OX14 4RQ, UK. [3]IBM Research UK, STFC Daresbury Laboratory, Warrington WA4 4AD, UK. [4]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK. [5]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. [6]Sistemic, West of Scotland Science Park, Glasgow G20 0SP, UK.

Received: 22 May 2018 Accepted: 31 August 2018
Published online: 20 October 2018

References

1. Laughlin CA, Morens DM, Cassetti MC, Costero-Saint Denis A, San Martin JL, Whitehead SS, et al. Dengue research opportunities in the Americas. J Infect Dis. 2012;206:1121–7.
2. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013;496:504–7.
3. Musso D, Cao-Lormeau VM, Gubler DJ. Zika virus: following the path of dengue and chikungunya? Lancet. 2015;386:243–4.
4. Fauci AS, Morens DM. Zika virus in the Americas - yet another arbovirus threat. N Engl J Med. 2016;374:601–4.
5. Alphey L. Genetic control of mosquitoes. Annu Rev Entomol. 2014;59:205–24.
6. World Health Organization. Mosquito (vector) control emergency response and preparedness for Zika virus. 2016. http://www.who.int/neglected_diseases/news/mosquito_vector_control_response/en/ Accessed 25 Apr 2016.
7. Gilles JRL, Schetelig MF, Scolari F, Marec F, Capurro ML, Franz G, et al. Towards mosquito sterile insect technique programmes: exploring genetic, molecular, mechanical and behavioural methods of sex separation in mosquitoes. Acta Trop. 2014;132:S178–87.
8. Hoang KP, Teo TM, Ho TX, Le VS. Mechanisms of sex determination and transmission ratio distortion in *Aedes aegypti*. Parasit Vectors. 2016;9:49.
9. Charlesworth D, Mank JE. The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. Genetics. 2010;186:9–31.
10. Craig GB, Hickey WA, Vandehey RC. An inherited male-producing factor in *Aedes aegypti*. Science. 1960;132:1887–9.
11. Clements AN. The Biology of Mosquitoes. London: Chapman & Hall; 1992.
12. Newton ME, Wood RJ, Southern DI. Cytological mapping of the M and D loci in the mosquito, *Aedes aegypti* (L.). Genetica. 1978;48:137–43.

13. Toups MA, Hahn MW. Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. Genetics. 2010;186:763–6.
14. Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler JK, et al. A male-determining factor in the mosquito *Aedes aegypti*. Science. 2015;348:1268–70.
15. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of *Aedes aegypti*. a major arbovirus vector. Science. 2007;316: 1718–23.
16. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015;23:110–20.
17. Severson DW, Behura SK. Mosquito genomics: progress and challenges. Annu Rev Entomol. 2012;57:143–66.
18. Hall AB, Timoshevskiy VA, Sharakhova MV, Jiang X, Basu S, Anderson MAE, et al. Insights into the preservation of the homomorphic sex-determining chromosome of *Aedes aegypti* from the discovery of a male-biased gene tightly linked to the M-locus. Genome Biol Evol. 2014;6:179–91.
19. Adelman ZN, Tu Z. Control of mosquito-borne infectious diseases: sex and gene drive. Trends Parasitol. 2016;32:219–29.
20. Matthews BJ, McBride CS, DeGennaro M, Despo O, Vosshall LB. The neurotranscriptome of the *Aedes aegypti* mosquito. BMC Genomics. 2016;17:32.
21. Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci USA. 2015;112:E5907–15.
22. Matthews BJ, Dudchenko O, Kingan S, Koren S, Antoshechkin I, Crawford JE, et al. Improved *Aedes aegypti* mosquito reference genome assembly enables biological discovery and vector control. bioRxiv. 2017;240747
23. Biedler JK, Hu W, Tae H, Tu Z. Identification of early zygotic genes in the yellow fever mosquito *Aedes aegypti* and discovery of a motif involved in early zygotic genome activation. PLoS One. 2012;7:e33933.
24. Artieri CG, Fraser HB. Transcript length mediates developmental timing of gene expression across *Drosophila*. Mol Biol Evol. 2014;31:2879–89.
25. Muller HJ. The relation of recombination to mutational advance. Mutat Res. 1964;1:2–9.
26. Charlesworth B. Evolution of sex chromosomes. Science. 1991;251:1030–3.
27. Kaiser VB, Bachtrog D. Evolution of sex chromosomes in insects. Annu Rev Genet. 2010;44:91–112.
28. Hall AB, Papathanos P-A, Sharma A, Cheng C, Akbari OS, Assour L, et al. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. Proc Natl Acad Sci USA. 2016;113:E2114–23.
29. Marais GAB, Nicolas M, Bergero R, Chambrier P, Kejnovsky E, Monéger F, et al. Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. Curr Biol. 2008;18:545–9.
30. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. 2013;14:113–24.
31. Carvalho AB, Dobo BA, Vibranovski MD, Clark AG. Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. Proc Natl Acad Sci USA. 2001;98:13225–30.
32. Charlesworth D, Charlesworth B, Marais G. Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb). 2005;95:118–28.
33. Fontaine A, Filipović I, Fansiri T, Hoffmann AA, Cheng C, Kirkpatrick M, et al. Extensive genetic differentiation between homomorphic sex chromosomes in the mosquito vector, *Aedes aegypti*. Genome Biol Evol. 2017;9:2322–35.
34. Lacroix R, McKemey AR, Raduan N, Kwee Wee L, Hong Ming W, Guat Ney T, et al. Open field release of genetically engineered sterile male *Aedes aegypti* in Malaysia. PLoS One. 2012;7:e42771.
35. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33:623–30.

ARTICLE

OPEN

https://doi.org/10.1038/s41586-018-0692-z

Improved reference genome of *Aedes aegypti* informs arbovirus vector control

Benjamin J. Matthews[1,2,3,49]*, Olga Dudchenko[4,5,6,7,49], Sarah B. Kingan[8,49], Sergey Koren[9], Igor Antoshechkin[10], Jacob E. Crawford[11], William J. Glassford[12], Margaret Herre[1,3], Seth N. Redmond[13,14], Noah H. Rose[15,16], Gareth D. Weedall[17,18], Yang Wu[19,20,21], Sanjit S. Batra[4,5,6], Carlos A. Brito-Sierra[22,23], Steven D. Buckingham[24], Corey L. Campbell[25], Saki Chan[26], Eric Cox[27], Benjamin R. Evans[28], Thanyalak Fansiri[29], Igor Filipović[30], Albin Fontaine[31,32,33,34], Andrea Gloria-Soria[28,35], Richard Hall[8], Vinita S. Joardar[27], Andrew K. Jones[36], Raissa G. G. Kay[37], Vamsi K. Kodali[27], Joyce Lee[26], Gareth J. Lycett[17], Sara N. Mitchell[11], Jill Muehling[8], Michael R. Murphy[27], Arina D. Omer[4,5,6], Frederick A. Partridge[24], Paul Peluso[8], Aviva Presser Aiden[4,5,38,39], Vidya Ramasamy[36], Gordana Rašić[30], Sourav Roy[40], Karla Saavedra-Rodriguez[25], Shruti Sharan[22,23], Atashi Sharma[21,41], Melissa Laird Smith[8], Joe Turner[42], Allison M. Weakley[11], Zhilei Zhao[15,16], Omar S. Akbari[43,44], William C. Black IV[25], Han Cao[26], Alistair C. Darby[42], Catherine A. Hill[22,23], J. Spencer Johnston[45], Terence D. Murphy[27], Alexander S. Raikhel[40], David B. Sattelle[24], Igor V. Sharakhov[21,41,46], Bradley J. White[11], Li Zhao[47], Erez Lieberman Aiden[4,5,6,7,13], Richard S. Mann[12], Louis Lambrechts[31,33], Jeffrey R. Powell[28], Maria V. Sharakhova[21,41,46], Zhijian Tu[20,21], Hugh M. Robertson[48], Carolyn S. McBride[15,16], Alex R. Hastie[26], Jonas Korlach[8], Daniel E. Neafsey[13,14], Adam M. Phillippy[9] & Leslie B. Vosshall[1,2,3]

Female *Aedes aegypti* mosquitoes infect more than 400 million people each year with dangerous viral pathogens including dengue, yellow fever, Zika and chikungunya. Progress in understanding the biology of mosquitoes and developing the tools to fight them has been slowed by the lack of a high-quality genome assembly. Here we combine diverse technologies to produce the markedly improved, fully re-annotated AaegL5 genome assembly, and demonstrate how it accelerates mosquito science. We anchored physical and cytogenetic maps, doubled the number of known chemosensory ionotropic receptors that guide mosquitoes to human hosts and egg-laying sites, provided further insight into the size and composition of the sex-determining M locus, and revealed copy-number variation among glutathione *S*-transferase genes that are important for insecticide resistance. Using high-resolution quantitative trait locus and population genomic analyses, we mapped new candidates for dengue vector competence and insecticide resistance. AaegL5 will catalyse new biological insights and intervention strategies to fight this deadly disease vector.

An accurate and complete genome assembly is required to understand the unique aspects of mosquito biology and to develop control strategies to reduce their capacity to spread pathogens[1]. The *Ae. aegypti* genome is large (approximately 1.25 Gb) and highly repetitive, and a 2007 genome project (AaegL3)[2] was unable to produce a contiguous genome fully anchored to a physical chromosome map[3] (Fig. 1a). A more recent assembly, AaegL4[4], produced chromosome-length scaffolds that made it possible to detect larger-scale syntenic genomic regions in other species but suffered from short contigs (contig N50, 84 kb, meaning that half of the assembly is found on contigs >84 kb) and a correspondingly large number of gaps (31,018; Fig. 1b). Taking advantage of rapid advances in sequencing and assembly technology in the last decade, we used long-read Pacific Biosciences sequencing and Hi-C (a high-throughput sequencing method based on chromosome conformation capture) scaffolding to produce a new reference genome (AaegL5) that is highly contiguous, with a decrease of

[1]Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, NY, USA. [2]Howard Hughes Medical Institute, New York, NY, USA. [3]Kavli Neural Systems Institute, New York, NY, USA. [4]The Center for Genome Architecture, Baylor College of Medicine, Houston, TX, USA. [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [6]Department of Computer Science, Rice University, Houston, TX, USA. [7]Center for Theoretical and Biological Physics, Rice University, Houston, TX, USA. [8]Pacific Biosciences, Menlo Park, CA, USA. [9]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [10]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [11]Verily Life Sciences, South San Francisco, CA, USA. [12]Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [13]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [14]Department of Immunology and Infectious Disease, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [15]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA. [16]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [17]Vector Biology Department, Liverpool School of Tropical Medicine, Liverpool, UK. [18]Liverpool John Moores University, Liverpool, UK. [19]Department of Pathogen Biology, School of Public Health, Southern Medical University, Guangzhou, China. [20]Department of Biochemistry, Virginia Tech, Blacksburg, VA, USA. [21]Fralin Life Science Institute, Virginia Tech, Blacksburg, VA, USA. [22]Department of Entomology, Purdue University, West Lafayette, IN, USA. [23]Purdue Institute for Inflammation, Immunology and Infectious Disease, Purdue University, West Lafayette, IN, USA. [24]Centre for Respiratory Biology, UCL Respiratory, University College London, London, UK. [25]Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO, USA. [26]Bionano Genomics, San Diego, CA, USA. [27]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [28]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. [29]Vector Biology and Control Section, Department of Entomology, Armed Forces Research Institute of Medical Sciences (AFRIMS), Bangkok, Thailand. [30]Mosquito Control Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [31]Insect-Virus Interactions Group, Department of Genomes and Genetics, Institut Pasteur, Paris, France. [32]Unité de Parasitologie et Entomologie, Département des Maladies Infectieuses, Institut de Recherche Biomédicale des Armées, Marseille, France. [33]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 2000, Paris, France. [34]Aix Marseille Université, IRD, AP-HM, SSA, UMR Vecteurs – Infections Tropicales et Méditerranéennes (VITROME), IHU - Méditerranée Infection, Marseille, France. [35]The Connecticut Agricultural Experiment Station, New Haven, CT, USA. [36]Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK. [37]Department of Entomology, University of California Riverside, Riverside, CA, USA. [38]Department of Bioengineering, Rice University, Houston, TX, USA. [39]Department of Pediatrics, Texas Children's Hospital, Houston, TX, USA. [40]Department of Entomology, Center for Disease Vector Research and Institute for Integrative Genome Biology, University of California, Riverside, CA, USA. [41]Department of Entomology, Virginia Tech, Blacksburg, VA, USA. [42]Institute of Integrative Biology, University of Liverpool, Liverpool, UK. [43]Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. [44]Tata Institute for Genetics and Society, University of California, San Diego, La Jolla, CA, USA. [45]Department of Entomology, Texas A&M University, College Station, TX, USA. [46]Laboratory of Ecology, Genetics and Environmental Protection, Tomsk State University, Tomsk, Russia. [47]Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY, USA. [48]Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [49]These authors contributed equally: Benjamin J. Matthews, Olga Dudchenko, Sarah B. Kingan. *e-mail: bnmtthws@gmail.com

RESEARCH ARTICLE

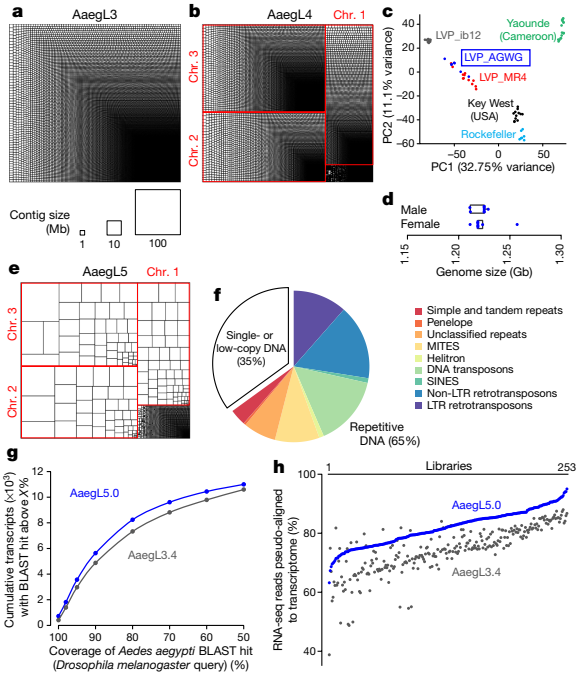


Fig. 1 | AaegL5 assembly statistics and annotation. a, **b**, Treemap of AaegL3 (**a**) and AaegL4 (**b**) contigs scaled by length. **c**, Principal component analysis of allelic variation of the indicated strains at 11,229 SNP loci. $n = 7$ per genotype **d**, Flow cytometry analysis of LVP_AGWG genome size. $n = 5$ per sex. Box plot: median is indicated by the blue line; boxes show first to third quartiles, whiskers are the 1.5× interquartile interval (Extended Data Fig. 1b). **e**, Treemap of AaegL5 contigs scaled by length. **f**, Genome composition (Supplementary Data 2, 3). **g**, Gene set alignment BLASTp coverage is compared between AaegL3.4 and AaegL5.0, with *D. melanogaster* protein queries. **h**, Alignment of 253 RNA-seq libraries to AaegL3.4 and AaegL5.0 gene set annotations (Supplementary Data 4–9). LTR, long terminal repeat retrotransposon; MITES, miniature inverted-repeat transposable elements; SINES, short interspersed nuclear elements.

93% in the number of contigs, and anchored end-to-end to the three *Ae. aegypti* chromosomes (Fig. 1 and Extended Data Figs. 1, 2). Using optical mapping and linked-read sequencing, we validated the local structure and predicted structural variants between haplotypes. We generated an improved gene set annotation (AaegL5.0), as assessed by a mean increase in RNA-sequencing (RNA-seq) read alignment of 12%, connections between many gene models that were previously split across multiple contigs, and a roughly twofold increase in the enrichment of assay for transposase-accessible chromatin using sequencing (ATAC-seq) alignments near predicted transcription start sites. We demonstrate the utility of AaegL5 and the AaegL5.0 annotation by investigating a number of scientific questions that could not be addressed with the previous genome annotations.

This project used the Liverpool *Aedes* Genome Working Group (LVP_AGWG) strain, related to the AaegL3 Liverpool ib12 (LVP_ib12) assembly strain[2] (Fig. 1c and Extended Data Fig. 1a). Using flow cytometry, we estimated that the genome size of LVP_AGWG is approximately 1.22 Gb (Fig. 1d and Extended Data Fig. 1b). To generate our primary assembly, we produced 166 Gb of Pacific Biosciences data (around 130× coverage for a 1.28-Gb genome) and assembled the genome using FALCON-Unzip[5]. This resulted in a total assembly length of 2.05 Gb (contig N50, 0.96 Mb; and NG50, 1.92 Mb, meaning half of the expected genome size found on contigs >1.92 Mb). FALCON-Unzip annotated the resulting contigs as either primary (3,967 contigs; N50, 1.30 Mb; NG50, 1.91 Mb) or haplotigs (3,823 contigs; N50, 193 kb), representing alternative haplotypes present in the approximately 80 male siblings pooled for sequencing (Table 1 and Extended Data Fig. 1e). The primary assembly was longer than expected for a haploid *Ae. aegypti* genome, as predicted by flow cytometry and prior assemblies, which was consistent with remaining alternative haplotypes that were too divergent to be automatically identified as primary and associated alternative haplotig pairs.

To generate a linear chromosome-scale reference genome assembly, we combined the primary contigs and haplotigs that were generated by FALCON-Unzip to create an assembly comprising 7,790 contigs. We used Hi-C to order and orient these contigs, correct misjoined sections and merge overlaps (Extended Data Fig. 1c–e). We set aside 359 contigs that were shorter than 20 kb and used the Hi-C data to identify 258 misjoined sections, resulting in 8,306 ordered and oriented contigs. This procedure revealed extensive sequence overlap among the contigs, consistent with the assembly of numerous alternative haplotypes. We developed a procedure to merge these alternative haplotypes, removing 5,440 gaps and boosting the contiguity (N50, 5.0 Mb; NG50, 4.6 Mb). This procedure placed 94% of sequenced (non-duplicated) bases onto three chromosome-length scaffolds that correspond to the three *Ae. aegypti* chromosomes. After scaffolding, we performed gap-filling and polishing using Pacific Biosciences reads. This removed 270 gaps and further increased the contiguity (N50, 11.8 Mb; NG50, 11.8 Mb), resulting in a final 1.279-Gb AaegL5 assembly and a complete mitochondrial genome (Fig. 1e and Table 1). We used Hi-C contact maps to estimate the position of the centromere with a resolution of around 5 Mb: chromosome 1, approximately 150–154 Mb; chromosome 2, around 227–232 Mb, chromosome 3, around 196–201 Mb. There are 229 remaining gaps in the primary assembly, including 173 on the three primary chromosomal scaffolds (Extended Data Fig. 2a and

Table 1 | Comparison of assembly statistics

	Genome assembly			
	AaegL3	AaegL4	AaegL5 FALCON-Unzip	AaegL5 (NCBI) FALCON-Unzip + Hi-C + polish
Total length (non-N bp)	1,310,092,987	1,254,548,160	1,695,064,654	1,278,709,169
Contig number	36,205	37,224	3,967	2,539
Contig N50 (bp)	82,618	84,074	1,304,397	11,758,062
Contig NG50 (bp)	85,043	81,911	1,907,936	11,758,062
Scaffold number	4,757	6,206	N/A	2,310
Scaffold N50 (bp)	1,547,048	404,248,146[a]	N/A	409,777,670[a]
GC content (%)	38.27	38.28	38.16	38.18
Alternative haplotypes (bp)	N/A	N/A	351,566,101	591,941,260
Alternative haplotypes (contigs)	N/A	N/A	3,823	4,224

N/A, not applicable.
[a]Scaffold N50 is the length of chromosome 3.

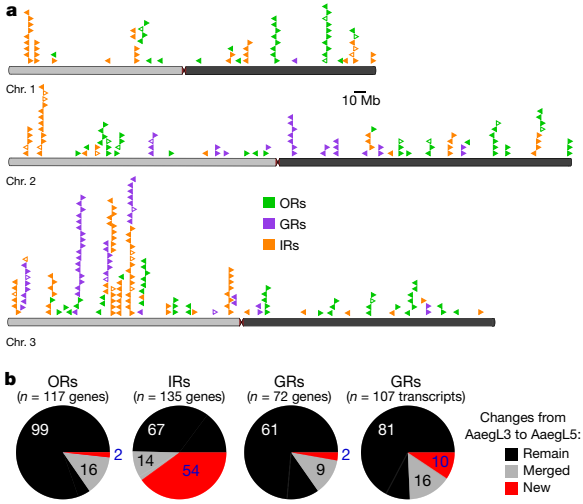


Fig. 2 | Chromosomal arrangement and increased number of chemosensory receptor genes. a, Location of predicted chemoreceptors (odorant receptors (ORs), gustatory receptors (GRs) and ionotropic receptors (IRs)) by chromosome in AaegL5. The blunt end of the arrowheads marks gene position and the arrow indicates orientation. Filled and open arrowheads represent intact genes and pseudogenes, respectively (Supplementary Data 17–20 and Extended Data Fig. 3). **b**, Chemosensory receptor annotations are compared between AaegL5 and AaegL3.

Supplementary Data 1). Analysis of near-universal single-copy orthologues using BUSCO[6] revealed a slight increase in complete single-copy orthologues and a reduction in fragmented and missing genes compared to previous assemblies (see Supplementary Methods and Supplementary Discussion). AaegL5 is markedly more contiguous than AaegL3 and AaegL4 assemblies[2,4] (Fig. 1a, b, e and Table 1). Using the TEfam, Repbase and de novo identified repeat databases, we found that 65% of AaegL5 was composed of transposable elements and other repetitive sequences (Fig. 1f and Supplementary Data 2, 3).

Complete and correct gene models are essential for studying all aspects of mosquito biology. We used the NCBI RefSeq annotation pipeline to produce annotation version 101 (AaegL5.0; Extended Data Fig. 2b) followed by manual curation of key gene families. AaegL5.0 formed the basis for a comprehensive quantification of transcript abundance in 253 sex-, tissue- and developmental stage-specific RNA-seq libraries (Supplementary Data 4–8). The AaegL5.0 gene set is considerably more complete and correct than previous versions. Many more genes have high protein coverage when compared to *Drosophila melanogaster* orthologues (915 more genes with >80% coverage, a 12.5% increase over AaegL3.4; Fig. 1g) and >12% more RNA-seq reads map to the AaegL5.0 gene set annotation than AaegL3.4 (Fig. 1h and Supplementary Data 9). In addition, 1,463 genes that were previously annotated separately as paralogues were collapsed into single gene models and 481 previously fragmented gene models were completed (Supplementary Data 10, 11). For example, *sex peptide receptor* is represented by a six-exon gene model in AaegL5.0 compared to two partial gene fragments on separate scaffolds in AaegL3.4 (Extended Data Fig. 2c). Genome-wide, we mapped a 1.8-fold higher number of ATAC-seq reads, known to co-localize with promoters and other *cis*-regulatory elements[7], to predicted transcription start sites in AaegL5.0 compared to AaegL3.4, consistent with more complete gene models in AaegL5.0 (Extended Data Fig. 2d).

We next validated the base-level and structural accuracy of the AaegL5 assembly. We estimate the lower bound of base-level accuracy of the assembly to have a quality value of 34.75 (meaning that 99.9665% of bases are correct, see Supplementary Methods and Supplementary Discussion). To develop a fine-scale physical genome map based on AaegL5, we compared the assembly coordinates of 500 bacterial artificial chromosome (BAC) clones containing *Ae. aegypti* genomic DNA with physical mapping by fluorescence in situ hybridization (FISH) (Extended Data Fig. 2e and Supplementary Data 12). After removing repetitive BAC-end sequences and those with ambiguous FISH signals, 377 out of 387 (97.4%) of probes showed concordance between physical mapping and BAC-end alignment. The 10 remaining discordant signals were not supported by Bionano or 10X analysis, and therefore probably do not reflect misassemblies in AaegL5. The genome coverage of this physical map is 93.5%, compared to 45% of AaegL3[8], and is one of the most complete genome maps across mosquito species[9,10].

Curation of multi-gene families

Large multi-gene families are very difficult to assemble and correctly annotate, because recently duplicated genes typically share high sequence similarities or can be misclassified as alleles of a single gene. We curated genes in large multi-gene families that encode proteases, G protein-coupled receptors, and chemosensory receptors using the improved AaegL5 genome and AaegL5.0 annotation. Serine proteases mediate immune responses[11] and metalloproteases have been linked to vector competence and mosquito–*Plasmodium* interactions[12]. Gene models for over 50% of the 404 annotated serine proteases and metalloproteases in AaegL3.4 were improved in AaegL5.0, and we found 49 previously unannotated protease genes (Supplementary Data 13). G protein-coupled receptors are membrane proteins that respond to diverse external and internal sensory stimuli. We provide major corrections to gene models that encode 10 visual opsins and 17 dopamine and serotonin receptors (Extended Data Fig. 2f and Supplementary Data 14–16). Three large multi-gene families of insect chemosensory receptors are ligand-gated ion channels: odorant receptors (*OR* gene family), gustatory receptors (*GR* gene family) and ionotropic receptors (*IR* gene family). These collectively allow insects to sense a vast array of chemical cues, including carbon dioxide and human body odours that activate and attract female mosquitoes. We identified 117 odorant receptors, 72 gustatory receptors (encoding 107 transcripts) and 135 ionotropic receptors in the AaegL5 assembly (Fig. 2a, b, Extended Data Fig. 3 and Supplementary Data 17–20), inferred new phylogenetic trees for each family to investigate the relationship of these receptors in *Ae. aegypti*, *Anopheles gambiae* malaria mosquitoes and *D. melanogaster* (Extended Data Figs. 4–6), and revised expression estimates for adult male and female neural tissues using deep RNA-seq[13] (Extended Data Fig. 7). Our annotation identified 54 new ionotropic receptor genes (Fig. 2b, Extended Data Fig. 3 and Supplementary Data 17), nearly doubling the known members of this family in *Ae. aegypti*. We additionally reannotated ionotropic receptors in *An. gambiae* and found 64 new genes. In *Ae. aegypti*, chemoreceptors are extensively clustered in tandem arrays (Fig. 2a and Extended Data Fig. 3), in particular on chromosome 3p, in which over a third of all chemoreceptor genes (n = 111) are found within a 109-Mb stretch. Although 71 gustatory receptor genes are scattered across chromosomes 2 and 3, only *AaegGr2*, a subunit of the carbon-dioxide receptor, is found on chromosome 1. Characterization of the full chemosensory receptor repertoire will enable the development of novel strategies to disrupt mosquito biting behaviour.

Structure of the sex-determining M locus

Sex determination in *Aedes* and *Culex* mosquitoes is governed by a dominant male-determining factor (M factor) at a male-determining locus (M locus) on chromosome 1[14–16]. This chromosome is homomorphic between the sexes except for the M/m karyotype, meaning that males are M/m and females are m/m. Despite the recent discovery of the M factor *Nix* in *Ae. aegypti*[17], which was entirely missing from the previous *Ae. aegypti* genome assemblies[2,4], the full molecular properties of the M locus remain unknown. We aligned AaegL5 (from M/m males) and AaegL4 (from m/m females), and identified a region that contained *Nix* in AaegL5 at which the assemblies diverged and that

RESEARCH ARTICLE

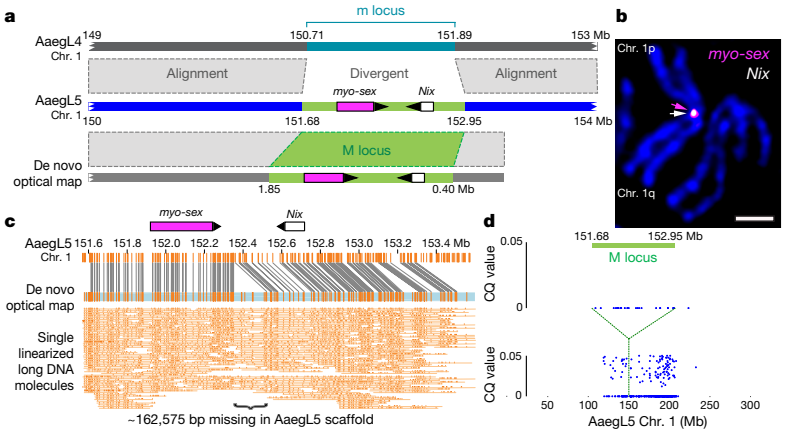


Fig. 3 | Application of AaegL5 to resolve the sex-determining locus. **a**, M locus structure indicating high alignment identity (grey-dashed boxes) and boundaries of *myo-sex* and *Nix* gene models (magenta and white boxes, arrowheads represent orientation). **b**, FISH of BAC clones containing *myo-sex* and *Nix*. Scale bar, 2 μm. Representative image of 10 samples. **c**, De novo optical map spanning the M locus and bridging the estimated 163-kb gap in the AaegL5 assembly. DNA molecules are cropped at the edges for clarity. **d**, Chromosome quotient (CQ) analysis of genomic DNA from male and female libraries aligned to AaegL5 chromosome 1. Each dot represents the CQ value of a repeat-masked 1-kb window with >20 reads aligned from male libraries.

may represent the divergent M/m locus (Fig. 3a). A de novo optical map assembly spanned the putative AaegL5 M locus and extended beyond its two borders. We estimated the size of the M locus at approximately 1.5 Mb, including an approximately 163-kb gap between contigs (Fig. 3a, c). We tentatively identified the female m locus as the region in AaegL4 not shared with the M locus-containing chromosome 1, but note that the complete phased structure of the divergent male M locus and corresponding female m locus remain to be determined. *Nix* contains a single intron of 100 kb, while *myo-sex*, a gene encoding a myosin heavy chain protein that has previously been shown to be tightly linked to the M locus[18], is approximately 300 kb in length. More than 73.7% of the M locus is repetitive: long terminal repeat retrotransposons comprise 29.9% of the M locus compared to 11.7% genome-wide. Chromosomal FISH with *Nix*- and *myo-sex*-containing BAC clones[19] showed that these genes co-localize to the 1p pericentromeric region (1p11) in only one homologous copy of chromosome 1, supporting the placement of the M locus at this position in AaegL5 (Fig. 3b). We investigated the differentiation between the sex chromosomes (Fig. 3d) using a chromosome quotient method to quantify regions of the genome with a strictly male-specific signal[20]. A sex-differentiated region in the LVP_AGWG strain extends to an approximately 100-Mb region surrounding the approximately 1.5-Mb M locus. This is consistent with the recent analysis of male–female F_{ST} in wild population samples and linkage map intercrosses[21] and could be explained by a large region of reduced recombination encompassing the centromere and M locus[22]. The availability of a more completely assembled mosquito M locus provides opportunities to study the evolution and maintenance of homomorphic sex-determining chromosomes. The sex-determining

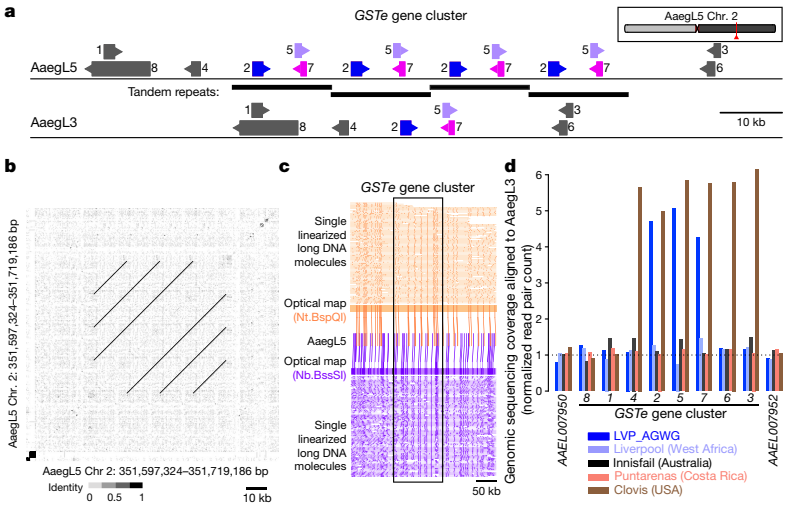


Fig. 4 | Copy-number variation in the glutathione *S*-transferase epsilon gene cluster. **a**, Glutathione *S*-transferase epsilon (*GSTe*) gene cluster structure in AaegL5 compared to AaegL3 (Supplementary Data 23). Arrowheads indicate gene orientation. **b**, Dot-plot alignment of AaegL5 *GSTe* region to itself. **c**, Optical mapping of DNA labelled with indicated enzymes. DNA molecules are cropped at the edges for clarity. **d**, Genomic sequencing coverage of AaegL3 *GSTe* genes (DNA read pairs mapped to each gene, normalized by gene length in kb) from one LVP_AGWG male and pooled mosquitoes from four other laboratory strains.

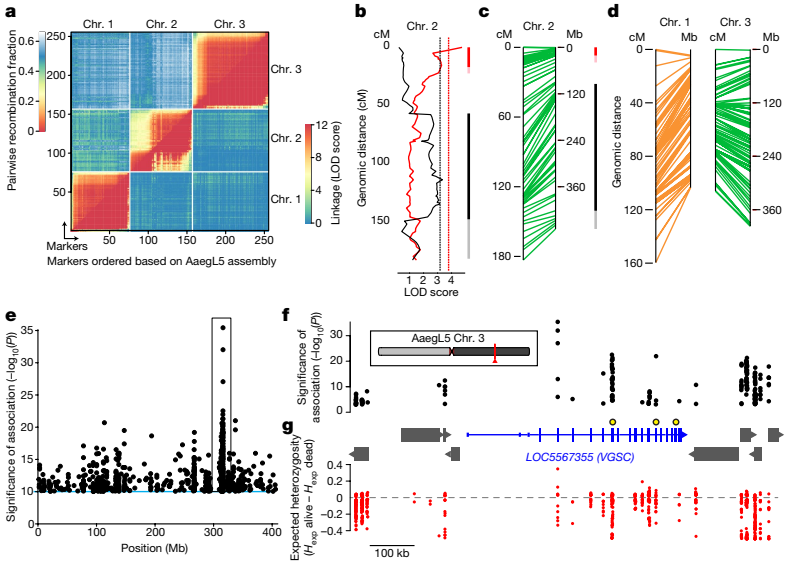


Fig. 5 | Using the AaegL5 genome for applied population genetics. **a**, Heat map of linkage based on pairwise recombination fractions for 255 RAD markers ordered by AaegL5 physical coordinates. **b**, Significant QTLs on chromosome 2 underly systemic DENV dissemination in midgut-infected mosquitoes (Extended Data Fig. 10a). Curves represent log of the odds ratio (LOD) scores obtained by interval mapping. Dotted vertical lines indicate genome-wide statistical significance thresholds ($\alpha = 0.05$). Confidence intervals of significant QTLs: bright colour, 1.5-LOD interval; light colour, 2-LOD interval with generalist effects (black, across DENV serotypes and isolates) and DENV isolate-specific effects (red, indicative of genotype-by-genotype interactions). **c**, **d**, Synteny between linkage map (in cM) and physical map (in Mb) for chromosome 2 (**c**) and chromosomes 1 and 3 (**d**). The orange color of chromosome 1 denotes uncertainty in the cM estimates because of deviations in Mendelian ratios surrounding the M locus. **e**, Chromosome 3 SNPs significantly correlated with deltamethrin survival. **f**, **g**, Magnified and inverted view of box in **e**, centred on the new gene model of voltage-gated sodium channel (*VGSC*, transcript variant X3; the chromosomal position is indicated in red). **f**, Non-coding genes are omitted for clarity, and other genes indicated with grey boxes. *VGSC* exons are represented by tall boxes and untranslated regions by short boxes. Arrowheads indicate gene orientation. Non-synonymous *VGSC* SNPs are marked with larger black and yellow circles: V1016I = 315,983,763; F1534C = 315,939,224; V410L = 316,080,722. **g**, Difference in expected heterozygosity (H_{exp} alive − H_{exp} dead) for all SNPs.

chromosome of *Ae. aegypti* may have remained homomorphic at least since the evolutionary divergence between the *Aedes* and *Culex* genera more than 50 million years ago. With the more completely assembled M locus, we can investigate how these chromosomes have avoided the proposed eventual progression into heteromorphic sex chromosomes[23].

Structural variation and gene families

Structural variation is associated with the capacity to vector pathogens[24]. We produced 'read cloud' Illumina sequencing libraries of linked reads with long-range (around 80 kb) phasing information from one male and one female mosquito using the 10X Genomics Chromium platform to investigate structural variants, including insertions, deletions, translocations and inversions, in individual mosquitoes. We observed abundant small-scale insertions and deletions (indels; 26 insertions and 81 deletions called, median 42.9 kb) and inversions and/or translocations (29 called) in these two individuals (Extended Data Fig. 8a and Supplementary Data 21). Eight of the inversions and translocations coincided with structural variants seen independently by Hi-C or FISH, suggesting that those variants are relatively common within this population and can be detected by different methods. AaegL5 will provide a foundation for the study of structural variants across *Ae. aegypti* populations.

Hox genes encode highly conserved transcription factors that specify segment identity along the anterior–posterior body axis of all metazoans[25]. In most vertebrates, *Hox* genes are clustered in a co-linear arrangement, although they are often disorganized or split in other animal lineages[26]. All expected *Hox* genes are present as a single copy in *Ae. aegypti*, but we identified a split between *labial* and *proboscipedia* placing *labial* on a separate chromosome (Extended Data Fig. 8b and Supplementary Data 22). We confirmed this in AaegL4, which was generated with Hi-C contact maps from a different *Ae. aegypti* strain[4], and note a similar arrangement in *Culex quinquefasciatus*, suggesting that it occurred before these two species diverged. Although this is not unprecedented[27], a unique feature of this organization is that both *labial* and *proboscipedia* appear to be close to telomeres.

Glutathione *S*-transferases (GSTs) are a large multi-gene family involved in the detoxification of compounds such as insecticides. Increased GST activity has been associated with resistance to multiple classes of insecticides, including organophosphates, pyrethroids and the organochlorine dichlorodiphenyltrichloroethane (DDT)[28]. Amplification of detoxification genes is one mechanism by which insects can develop insecticide resistance[29]. We found that three insect-specific GST epsilon (GSTe) genes on chromosome 2, located centrally in the cluster (*GSTe2*, *GSTe5* and *GSTe7*), are duplicated four times in AaegL5 relative to AaegL3 (Fig. 4a, b and Supplementary Data 23). Short Illumina read coverage and optical maps confirmed the copy number and arrangement of these duplications in AaegL5 (Fig. 4c, d), and analysis of whole-genome sequencing data for four additional laboratory colonies showed variable copy numbers across this gene cluster (Fig. 4d). GSTe2 is a highly efficient metaboliser of DDT[30], and it is noteworthy that the cDNA from three GST genes in the quadruplication was detected at higher levels in DDT-resistant *Ae. aegypti* mosquitoes from southeast Asia[31].

Genome-wide genetic variation

Measurement of genetic variation within and between populations is important for inferring ongoing and historic evolution in a species[32]. To understand genomic diversity in *Ae. aegypti*, which spread in the last century from Africa to tropical and subtropical regions around the world, we performed whole-genome resequencing on four laboratory colonies. Chromosomal patterns of nucleotide diversity should correlate with regional differences in meiotic recombination rates[33].

We observed pronounced declines in genetic diversity near the centre of each chromosome (Extended Data Fig. 9a, b), providing independent corroboration of the estimated position of each centromere by Hi-C (Extended Data Fig. 2a).

To investigate linkage disequilibrium in geographically diverse populations of *Ae. aegypti*, we first mapped Affymetrix SNP-Chip markers that were designed using AaegL3[34] to positions on AaegL5. We genotyped 28 individuals from two populations from Amacuzac, Mexico and Lopé National Park, Gabon and calculated the pairwise linkage disequilibrium of single-nucleotide polymorphisms (SNPs) from 1-kb bins both genome-wide and within each chromosome (Extended Data Fig. 9c, d). The maximum linkage disequilibrium in the Mexican population is approximately twice that of the population from Gabon, which probably reflects a recent bottleneck associated with the spread of this species out of Africa.

Dengue competence and pyrethroid resistance

To illustrate the value of AaegL5 for mapping quantitative trait loci (QTLs), we used restriction site-associated DNA (RAD) markers to locate QTLs underlying dengue virus (DENV) vector competence. We identified and genotyped RAD markers in the F_2 progeny of a laboratory cross between wild *Ae. aegypti* founders from Thailand[35] (Extended Data Fig. 10a). For this population, 197 F_2 females had previously been scored for DENV vector competence against four different DENV isolates (two isolates from serotype 1 and two from serotype 3)[35]. The newly developed linkage map included a total of 255 RAD markers (Fig. 5a) with perfect concordance between genetic distances in centiMorgans (cM) and AaegL5 physical coordinates in Mb (Fig. 5a, c, d). We detected two significant QTLs on chromosome 2 that underlie the likelihood of DENV dissemination from the midgut (that is, systemic infection), an important component of DENV vector competence[36]. One QTL was associated with a generalist effect across DENV serotypes and isolates, whereas the other was associated with an isolate-specific effect (Fig. 5b, c). QTL mapping powered by AaegL5 will make it possible to understand the genetic basis of *Ae. aegypti* vector competence for arboviruses.

Pyrethroid insecticides are used to combat mosquitoes, including *Ae. aegypti*, and emerging resistance to these compounds is a global problem[37]. Understanding the mechanisms that underlie insecticide targets and resistance in different mosquito populations is critical to combating arboviral pathogens. Many insecticides act on ion channels, and we curated members of the Cys-loop ligand-gated ion channel (Cys-loop LGIC) superfamily in AaegL5. We found 22 subunit-encoding Cys-loop LGICs (Extended Data Fig. 10d and Supplementary Data 24), of which 14 encode nicotinic acetylcholine receptor (nAChR) subunits. nAChRs consist of a core group of subunit-encoding genes ($\alpha1$–$\alpha8$ and $\beta1$) that are highly conserved between insect species, and at least one divergent subunit[38]. Whereas *D. melanogaster* possesses only one divergent nAChR subunit, *Ae. aegypti* has five. We found that agricultural and veterinary insecticides impaired the motility of *Ae. aegypti* larvae (Extended Data Fig. 10c), suggesting that these Cys-loop LGIC-targeting compounds have potential as mosquito larvicides. The improved annotation presented here provides a valuable resource for investigating insecticide efficacy.

To demonstrate how a chromosome-scale genome assembly informs genetic mechanisms of insecticide resistance, we performed a genome-wide population genetic screen for SNPs correlating with resistance to deltamethrin in *Ae. aegypti* collected in Yucatán, Mexico, where pyrethroid-resistant and -susceptible populations co-exist (Fig. 5e). We uncovered an association with non-synonymous changes to three amino acid residues of the voltage-gated sodium channel VGSC, a known target of pyrethroids (Fig. 5f). The gene model for *VGSC*, a complex locus spanning nearly 500 kb in AaegL5, was incomplete and highly fragmented in AaegL3. SNPs in this region have a lower expected heterozygosity (H_{exp}) in the resistant compared to the susceptible population, suggesting that they are part of a selective sweep for the resistance phenotype surrounding *VGSC* (Fig. 5g). Accurately associating SNPs with phenotypes requires a fully assembled genome, and we expect that AaegL5 will be critical to understanding the evolution of insecticide resistance and other important traits.

Summary

The high-quality genome assembly and annotation described here will enable major advances in mosquito biology, and has already allowed us to carry out a number of experiments that were previously impossible. The highly contiguous AaegL5 genome permitted high-resolution genome-wide analysis of genetic variation and the mapping of loci for DENV vector competence and insecticide resistance. A new appreciation of copy number variation in insecticide-detoxifying *GSTe* genes and a more complete accounting of Cys-loop LGICs will catalyse the search for new resistance-breaking insecticides. A doubling in the known number of chemosensory ionotropic receptors provides opportunities to link odorants and tastants on human skin to mosquito attraction, a key first step in the development of novel mosquito repellents. 'Sterile Insect Technique' and 'Incompatible Insect Technique' show great promise to suppress mosquito populations[39], but these population suppression methods require that only males are released. A strategy that connects a gene for male determination to a gene drive construct has been proposed to effectively bias the population towards males over multiple generations[40], and improved understanding of M locus evolution and the function of its genetic content should facilitate genetic control of mosquitoes that infect many hundreds of millions of people with arboviruses every year[1].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0692-z.

Received: 28 December 2017; Accepted: 5 October 2018;
Published online 14 November 2018.

1. Bhatt, S. et al. The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
2. Nene, V. et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316**, 1718–1723 (2007).
3. Timoshevskiy, V. A. et al. An integrated linkage, chromosome, and genome map for the yellow fever mosquito *Aedes aegypti*. *PLoS Negl. Trop. Dis.* **7**, e2052 (2013).
4. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
5. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
6. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2017).
7. Denny, S. K. et al. Nfib promotes metastasis through a widespread increase in chromatin accessibility. *Cell* **166**, 328–342 (2016).
8. Timoshevskiy, V. A. et al. Genomic composition and evolution of *Aedes aegypti* chromosomes revealed by the analysis of physically mapped supercontigs. *BMC Biol.* **12**, 27 (2014).
9. George, P., Sharakhova, M. V. & Sharakhov, I. V. High-resolution cytogenetic map for the African malaria vector *Anopheles gambiae*. *Insect Mol. Biol.* **19**, 675–682 (2010).
10. Artemov, G. N. et al. The physical genome mapping of *Anopheles albimanus* corrected scaffold misassemblies and identified interarm rearrangements in genus *Anopheles*. *G3 (Bethesda)* **7**, 155–164 (2017).
11. Gorman, M. J. & Paskewitz, S. M. Serine proteases as mediators of mosquito immune responses. *Insect Biochem. Mol. Biol.* **31**, 257–262 (2001).
12. Goulielmaki, E., Sidén-Kiamos, I. & Loukeris, T. G. Functional characterization of *Anopheles matrix metalloprotease 1* reveals its agonistic role during sporogonic development of malaria parasites. *Infect. Immun.* **82**, 4865–4877 (2014).
13. Matthews, B. J., McBride, C. S., DeGennaro, M., Despo, O. & Vosshall, L. B. The neurotranscriptome of the *Aedes aegypti* mosquito. *BMC Genomics* **17**, 32 (2016).
14. Gilchrist, B. M. & Haldane, J. B. S. Sex linkage and sex determination in a mosquito, *Culex molestus*. *Hereditas* **33**, 175–190 (1947).
15. McClelland, G. A. H. Sex-linkage in *Aedes aegypti*. *Trans. R. Soc. Trop. Med. Hyg.* **56**, 4 (1962).
16. Newton, M. E., Wood, R. J. & Southern, D. I. Cytological mapping of the M and D loci in the mosquito, *Aedes aegypti* (L.). *Genetica* **48**, 137–143 (1978).
17. Hall, A. B. et al. A male-determining factor in the mosquito *Aedes aegypti*. *Science* **348**, 1268–1270 (2015).
18. Hall, A. B. et al. Insights into the preservation of the homomorphic sex-determining chromosome of *Aedes aegypti* from the discovery of a male-biased gene tightly linked to the M-locus. *Genome Biol. Evol.* **6**, 179–191 (2014).

19. Turner, J. et al. The sequence of a male-specific genome region containing the sex determination switch in *Aedes aegypti*. *Parasit. Vectors* **11**, 549 (2018).
20. Hall, A. B. et al. Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females. *BMC Genomics* **14**, 273 (2013).
21. Fontaine, A. et al. Extensive genetic differentiation between homomorphic sex chromosomes in the mosquito vector, *Aedes aegypti*. *Genome Biol. Evol*. **9**, 2322–2335 (2017).
22. Juneja, P. et al. Assembly of the genome of the disease vector *Aedes aegypti* onto a genetic linkage map allows mapping of genes affecting disease transmission. *PLoS Negl. Trop. Dis.* **8**, e2652 (2014).
23. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118–128 (2005).
24. Riehle, M. M. et al. The *Anopheles gambiae* 2La chromosome inversion is associated with susceptibility to *Plasmodium falciparum* in Africa. *eLife* **6**, e25813 (2017).
25. Lewis, E. B. A gene complex controlling segmentation in *Drosophila*. *Nature* **276**, 565–570 (1978).
26. Duboule, D. The rise and fall of *Hox* gene clusters. *Development* **134**, 2549–2560 (2007).
27. Negre, B., Ranz, J. M., Casals, F., Cáceres, M. & Ruiz, A. A new split of the *Hox* gene complex in *Drosophila*: relocation and evolution of the gene *labial*. *Mol. Biol. Evol*. **20**, 2042–2054 (2003).
28. Enayati, A. A., Ranson, H. & Hemingway, J. Insect glutathione transferases and insecticide resistance. *Insect Mol. Biol*. **14**, 3–8 (2005).
29. Bass, C. & Field, L. M. Gene amplification and insecticide resistance. *Pest Manag. Sci*. **67**, 886–890 (2011).
30. Ortelli, F., Rossiter, L. C., Vontas, J., Ranson, H. & Hemingway, J. Heterologous expression of four glutathione transferase genes genetically linked to a major insecticide-resistance locus from the malaria vector *Anopheles gambiae*. *Biochem. J*. **373**, 957–963 (2003).
31. Lumjuan, N. et al. The role of the *Aedes aegypti* Epsilon glutathione transferases in conferring resistance to DDT and pyrethroid insecticides. *Insect Biochem. Mol. Biol*. **41**, 203–209 (2011).
32. Anopheles gambiae 1000 Genomes Consortium. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature* **552**, 96–100 (2017).
33. Begun, D. J. & Aquadro, C. F. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520 (1992).
34. Evans, B. R. et al. A multipurpose, high-throughput single-nucleotide polymorphism chip for the dengue and yellow fever mosquito, *Aedes aegypti*. *G3 (Bethesda)* **5**, 711–718 (2015).
35. Fansiri, T. et al. Genetic mapping of specific interactions between *Aedes aegypti* mosquitoes and dengue viruses. *PLoS Genet*. **9**, e1003621 (2013).
36. Black, W. C. IV et al. Flavivirus susceptibility in *Aedes aegypti*. *Arch. Med. Res*. **33**, 379–388 (2002).
37. Moyes, C. L. et al. Contemporary status of insecticide resistance in the major *Aedes* vectors of arboviruses infecting humans. *PLoS Negl. Trop. Dis*. **11**, e0005625 (2017).
38. Jones, A. K. & Sattelle, D. B. Diversity of insect nicotinic acetylcholine receptor subunits. *Adv. Exp. Med. Biol*. **683**, 25–43 (2010).
39. Alphey, L. Genetic control of mosquitoes. *Annu. Rev. Entomol*. **59**, 205–224 (2014).
40. Adelman, Z. N. & Tu, Z. Control of mosquito-borne infectious disease: sex and gene drive. *Trends Parasitol*. **32**, 219–229 (2016).

Acknowledgements We thank R. Andino; S. Emrich and D. Lawson (Vectorbase); A. A. James, M. Kunitomi, C. Nusbaum, D. Severson, N. Whiteman; T. Dickinson, M. Hartley and B. Rice (Dovetail Genomics) for early participation in the AGWG; C. Bargmann, D. Botstein, E. Jarvis and E. Lander for encouragement and facilitation. N. Keivanfar, D. Jaffe and D. M. Church (10X Genomics) prepared DNA for structural-variant analysis. We thank A. Harmon of the New York Times and acknowledge generous pro bono data and analysis from our corporate collaborators. This research was supported in part by federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under grant number U19AI110818 to the Broad Institute (S.N.R. and D.E.N.); USDA 2017-05741 (E.L.A.); NSF PHY-1427654 Center for Theoretical Biological Physics (E.L.A.); NIH Intramural Research Program, National Library of Medicine and National Human Genome Research Institute (A.M.P. and S.K.) and the following extramural NIH grants: R01AI101112 (J.R.P.), R35GM118336 (R.S.M. and W.J.G.), R21AI121853 (M.V.S., I.V.S. and A.S.), R01AI123338 (Z.T.), T32GM007739 (M.H.), NIH/NCATS UL1TR000043 (Rockefeller University), DP2OD008540 (E.L.A.), U01AI088647, 1R01AI121211 (W.C.B. IV), Fogarty Training Grant D43TW001130-08, U01HL130010 (E.L.A.), UM1HG009375 (E.L.A), 5K22AI113060 (O.S.A.), 1R21AI123937 (O.S.A.), and R0DC012069 (C.S.M.); Defence Advanced Research Project Agency: HR0011-17-2- 0047 (O.S.A.). Other support was provided by Jane Coffin Childs Memorial Fund (B.J.M.), Center for Theoretical Biological Physics postdoctoral fellowship (O.D.), Robertson Foundation (L.Z.), and McNair & Welch (Q-1866) Foundations (E.L.A.), French Government's Investissement d'Avenir program, Laboratoire d'Excellence Integrative Biology of Emerging Infectious Diseases (grant ANR-10-LABX-62-IBEID to L.L.), Agence Nationale de la Recherche grant ANR-17-ERC2-0016-01 (L.L.), European Union's Horizon 2020 research and innovation program under ZikaPLAN grant agreement no. 734584 (L.L.), Pew and Searle Scholars Programs (C.S.M.), Klingenstein-Simons Fellowship in the Neurosciences (C.S.M.). A.M.W., B.J.W., J.E.C. and S.N.M. were supported by Verily Life Sciences. L.B.V. is an investigator of the Howard Hughes Medical Institute.

Reviewer information *Nature* thanks S. Celniker, A. G. Clark, R. Waterhouse and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.J.M. and L.B.V. conceived the study, coordinated data collection and analysis, designed the figures and wrote the paper with input from all authors. B.J.M. developed and distributed animals and/or DNA of the LVP_AGWG strain. P.P., M.L.S. and J.M. carried out Pacific Biosciences sample preparation and sequencing. S.B.K., R.H., J.K., S.K. and A.M.P. were involved in genome assembly. A.R.H., S.C., J.L. and H.C. carried out Bionano optical mapping. O.D., S.S.B., A.D.O., A.P.A. and E.L.A. carried out Hi-C sample preparation, scaffolding and deduplication. The following authors contributed analysis and data to the indicated figures: B.R.E., A.G.-S. and J.R.P. (Fig. 1c); J.S.J. (Fig. 1d); L.Z. (Fig. 1f); E.C., V.S.J., V.K.K., M.R.M., T.D.M. and B.J.M. (Fig. 1g); I.A., O.S.A., J.E.C., A.M.W., B.J.W., R.G.G.K., S.N.M. and B.J.M. (Fig. 1h); C.S.M., H.M.R., Z.Z., N.H.R. and B.J.M. (Fig. 2); Z.T., M.V.S., I.V.S., A.S., Y.W., J.T., A.C.D., A.R.H. and B.J.M. (Fig. 3); G.D.W., B.J.M., A.R.H., S.B.K., A.M.P. and S.K. (Fig. 4); A.F., I.F., T.F., G.R. and L.L. (Fig. 5a–d); C.L.C., K.S.-R., W.C.B. and B.J.M. (Fig. 5e–g); B.J.M. (Extended Data Fig. 1a); J.S.J. (Extended Data Fig. 1b); O.D., S.S.B., A.D.O., A.P.A. and E.L.A. (Extended Data Fig. 1c, d); S.B.K., J.K., O.D., E.L.A., S.K., A.M.P. and B.J.M. (Extended Data Fig. 1e); A.R.H. and B.J.M. (Extended Data Fig. 2a); E.C., V.S.J., V.K.K., M.R.M., T.D.M. and B.J.M. (Extended Data Fig. 2b); M.H. and B.J.M. (Extended Data Fig. 2c, d); A.S., I.V.S. and M.V.S. (Extended Data Fig. 2e); C.A.B.-S., S.S. and C.A.H. (Extended Data Fig. 2f); C.S.M., H.M.R., Z.Z., N.H.R. and B.J.M. (Extended Data Figs. 3–7); ; S.N.R. and D.E.N. (Extended Data Fig. 8a); W.J.G., R.S.M., O.D., E.L.A. and B.J.M. (Extended Data Fig. 8b, c); W.J.G. and R.S.M. (Extended Data Fig. 8d); J.E.C., A.M.W., B.J.W., R.G.G.K. and S.N.M. (Extended Data Fig. 9a, b); B.R.E., A.G.-S. and J.R.P. (Extended Data Fig 9c, d); A.F., I.F., T.F, G.R. and L.L. (Extended Data Fig. 10a, b); G.J.L., A.K.J., V.R., S.D.B., F.A.P. and D.B.S. (Extended Data Fig. 10c, d); A.R.H. (Supplementary Data 1); L.Z. (Supplementary Data 2, 3); I.A., O.S.A., J.E.C., A.M.W., B.J.W., R.G.G.K., S.N.M. and B.J.M. (Supplementary Data 4–9); E.C., V.S.J., V.K.K., M.R.M. and T.D.M. (Supplementary Data 10, 11); A.S., I.V.S. and M.V.S. (Supplementary Data 12); S.R. and A.S.R. (Supplementary Data 13); C.A.B.-S., S.S. and C.A.H. (Supplementary Data 14–16); C.S.M., H.M.R., Z.Z., N.H.R. and B.J.M. (Supplementary Data 17–20); S.N.R. and D.E.N. (Supplementary Data 21); W.J.G. and R.S.M. (Supplementary Data 22); G.D.W. and B.J.M. (Supplementary Data 23); G.J.L., A.K.J., V.R., S.D.B., F.A.P. and D.B.S. (Supplementary Data 24).

Competing interests P.P., M.L.S., J.M., S.B.K., R.H. and J.K. are employees of Pacific Biosciences, a company developing single-molecule sequencing technologies. J.L., S.C., H.C. and A.R.H. are employees of Bionano Genomics and own company stock options. O.D., S.S.B., A.D.O., A.P.A. and E.L.A. are inventors on a US provisional patent application 62/347,605, filed 8 June 2016, by the Baylor College of Medicine and the Broad Institute.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0692-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0692-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.J.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix 2 Supplementary Information

2.1 Sequencing data downloaded

SRA Accession	Dataset	Sex	Instrument	Notes
SRR871496	Virginia Tech (LVP)	Female	Illumina HiSeq 2000	Whole genome sequencing
SRR871497	Virginia Tech (LVP)	Female	Illumina HiSeq 2000	Whole genome sequencing
SRR871499	Virginia Tech (LVP)	Male	Illumina HiSeq 2000	Whole genome sequencing
SRR871500	Virginia Tech (LVP)	Male	Illumina HiSeq 2000	Whole genome sequencing
SRR6063610	Rockefeller (LVP_AGWG)	Male	Illumina NextSeq 500	Whole genome sequencing
SRR6063611	Rockefeller (LVP_AGWG)	Female	Illumina NextSeq 500	Whole genome sequencing
SRR6063612	Rockefeller (LVP_AGWG)	Male	Illumina NextSeq 500	Whole genome sequencing
SRR6063613	Rockefeller (LVP_AGWG)	Female	Illumina NextSeq 500	Whole genome sequencing
SRR4868127	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868128	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868129	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868130	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868131	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868132	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868133	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868134	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868135	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868136	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868137	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868138	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868139	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868140	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868141	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868142	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868143	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868144	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868145	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868146	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868147	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868148	Cambridge	Male	Illumina HiSeq 2000	Exome sequencing
SRR4868149	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868150	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868151	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868152	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868153	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868154	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868155	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868156	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868157	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868158	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868159	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868160	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868161	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868162	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868163	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868164	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868165	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868166	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868167	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868168	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868169	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing

SRR4868170	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868171	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868172	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868173	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868174	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868175	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868176	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868177	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868178	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868179	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868180	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868181	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868182	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868183	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868184	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868185	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868186	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868187	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868188	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868189	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868190	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868191	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868192	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868193	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868194	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868195	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868196	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868197	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868198	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868199	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868200	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868201	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868202	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868203	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868204	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868205	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868206	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868207	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868208	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868209	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868210	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868211	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868212	Cambridge	Male	Illumina HiSeq 2000	Exome sequencing
SRR4868213	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868214	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868215	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868216	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868217	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868218	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868219	Cambridge	Male	Illumina HiSeq 2000	Exome sequencing
SRR4868220	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868221	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868222	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868223	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868224	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868225	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868226	Cambridge	Male	Illumina HiSeq 2000	Exome sequencing
SRR4868227	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868228	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing

SRR4868229	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868230	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868231	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868232	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868233	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868234	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868235	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868236	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868237	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868238	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868239	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868240	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868241	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868242	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868243	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868244	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868245	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868246	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868247	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868248	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868249	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868250	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868251	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868252	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868253	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868254	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868255	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868256	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868257	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868258	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868259	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868260	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868261	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868262	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868263	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868264	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868265	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868266	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868267	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR4868268	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868269	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868270	Cambridge	Missing	Illumina HiSeq 2000	Exome sequencing
SRR4868271	Cambridge	Female	Illumina HiSeq 2000	Exome sequencing
SRR1585314	Virginia Tech RNA-Seq (LVP)	Female	Illumina HiSeq 2000	RNA sequencing (whole adult)
SRR1585315	Virginia Tech RNA-Seq (LVP)	Female	Illumina HiSeq 2000	RNA sequencing (whole adult)
SRR1585316	Virginia Tech RNA-Seq (LVP)	Female	Illumina HiSeq 2000	RNA sequencing (whole adult)
SRR1585317	Virginia Tech RNA-Seq (LVP)	Male	Illumina HiSeq 2000	RNA sequencing (whole adult)
SRR1585318	Virginia Tech RNA-Seq (LVP)	Male	Illumina HiSeq 2000	RNA sequencing (whole adult)
SRR1585319	Virginia Tech RNA-Seq (LVP)	Male	Illumina HiSeq 2000	RNA sequencing (whole adult)
SRR924021	Caltech RNA-Seq (LVP)	Male	Illumina Genome Analyzer II	RNA sequencing (testes)
SRR5961503	Cambridge miRNA-Seq	Female	Illumina NextSeq 500	smRNA sequencing (thorax) (oxidation-treated)
SRR5961504	Cambridge miRNA-Seq	Female	Illumina NextSeq 500	smRNA sequencing (germline) (oxidation-treated)
SRR5961505	Cambridge miRNA-Seq	Female	Illumina NextSeq 500	smRNA sequencing (thorax)
SRR5961506	Cambridge miRNA-Seq	Female	Illumina NextSeq 500	smRNA sequencing (germline)
SRR1068553	Virginia Tech miRNA-Seq	Male	Illumina Genome Analyzer	smRNA sequencing (whole adult)

2.2 Primers used

Sequence_description	F_primer	F_primer_seq	R_primer	R_primer_seq
Common sgRNA sequence			SS1713	AAAAGCACCGACTCGGTGCCACTT TTTCAAGTTGATAACGGACTAGCC TTATTTTAACTTGCTATTTCTAGC TCTAAAAC
cas9	SS1765	AGCTGGTGCAGACCTACAACCAG	SS1853	CGGTTATCCTTCAGGAACGG
ku70	SS2081	TAATACGACTCACTATAGGGTTT CGCTGGGTGTCAACATTTATTCC GGAAAAAAAAACTTGATTTTCTT CGCTACTGACTTGA	SS2082	TAATACGACTCACTATAGGGCATG AGAAACAAGATCATCAGGGAAAAA AAGATCTTCGTCGGCTACCGTACA
Nix	Nix1F	TTGAGTCTGAAAAGTCTATGCAA	Nix1R	TCGCTCTTCCGTGGCATTTGA
Nix	Nix2F	ACGTAGTCGGCAACTCGAAG	Nix2R	CTGGGACAAATCGAACGGAA
Nix region Unique 5' end sequence	JT1	TTGACCCGGCCTTTCCTATT	JT2	CAAGCCACGTTCGATGAGAG
Nix region Unique putative genic region exon 1	JT3	AAGTGCAACCAGGTTATGCG	JT4	ATTGCTGCTGGTACTGTTGC
Nix region Unique putative genic region exon 2	JT5	GGAAAACCGAACAACCACGA	JT6	ATCTCGTCGTGGTTATCGCT
Nix region Unique putative genic region exon 3	JT7	ATGGCAGTGCAACAGTTCAG	JT8	TACTCTTGTCCGTCGCTTGT
Nix region 3' end sequence	JT9	ACCGGTCCTTGAACCTTTGTT	JT10	CCGTAAAGGCATTCCGTGTTC
Nix region 3' end sequence	JT11	ACCGGTCCTTGAACCTTTGT	JT12	TCTCTACAACCTGCGTTTGCT
Nix region 3' end sequence	JT13	CGTCACGTTGATCCCACAGA	JT14	AACGCAGCTCTCTACTTCCG
Nix region 5' end sequence reverse strand	JT15	AATCGCTAATGACACCGGCA	JT16	ACCGTTGATAAGCTGGCTCC
Nix region 5' end sequence reverse strand	JT17	TGTCGGAGGATGGTGGTACA	JT18	GCGCCAAACCTGTCTCAATC
Nix region 5' end sequence reverse strand	JT19	TTCGCTGGTTTCCAGGAGTG	JT20	GTTCTCTTCGAACGCCGTGG
Nix region Unique putative genic region intron 3	JT27	TTCTCCATACACACAGGCGT	JT28	GGCTAGGACCGTTAACCCTT
Nix region 5' end sequence	JT29	GAGTAATCGCCCTCGTCCAG	JT30	TTCAATGCGGAGAGGCACTT
Nix region 5' end sequence	JT31	TTGAAGCGACCATGTGACGA	JT32	AACTCGTCGGGTGAAAGCAT
Nix region 5' end sequence	JT33	CAGAGCTCGTACGCTCAGTT	JT34	TCGGGACAAGGTTGAATCGG
Nix region 5' end sequence	JT35	ACTGAGCTCACTTGCTCCAC	JT36	CGTCACCGGATATTACGCGA
Nix region 5' end sequence	JT37	TTTGTACAGGCTGACTGCGA	JT38	CCAACAGCGGAGATGCGATA
Myo-sex	JT39	CCTTCAAGCACACCGTTACA	JT40	TCACTATGCAGGAGTTGTTTCG
Myo-sex	JT41	GATGCAACGCACAAAATGAG	JT42	TGTGCAGTGTATTTTTCCCTGA

Myo-sex	JT43	AGGATGCTCGAAACCAGCTA	JT44	TGAGCAATTTCGTCGTTCAG
BAC NDL62N22	JT45	TGCATGAATCTGCTTGGGTA	JT46	AGGCGTGTAGACACGAAGGT
BAC NDL62N23	JT47	CAGTGGTTGGAAAGCACGTA	JT48	ATTTCCGAAGTGTGGAGCTG
AY988440.1 18S rDNA	JT49	CAACGGGTAACGGGGAATCA	JT50	GGTAATTTACGCGCCTGCTG
AY988440.1 18S rDNA	JT51	AACCAATCGTCCTCCGTGAC	JT52	TGATTCCCCGTTACCCGTTG
AY988440.1 18S rDNA	JT53	ACCACATCCAAGGAAGGCAG	JT54	CAAATTAAGCCGCAGGCTCC
AAGE02035037.1(1-6260)	M37	CGTGAAATCTATGACACAAGGTCG	M38	GCGAAATTCGGCGAGTGAAG
AAGE02035965.1(1-4650)	M41	GGTTGAGAAACGTCAGCTCTCTGA	M42	GCATGGCATGTGGTCAAGTTATC
AAGE02035016.1(1-6296)	M7	GTGCAGCGTGTGACTCACGTAC	M8	CGGTGAGTACTTCAGCGCAGAAG
AAGE02035557.1(1-5425)	M11	CCATGTCTGGTCATAACAGCAGAG	M12	CCTCTACATGACAAATTGAATGAAGC
AAGE02036067.1(1-4250)	SS1818	GTATAACCAACAACAATTCCACCG	SS1819	GAGCTTCTGGACTTCTGGACGA
AAGE02035994.1(1-4545)	M1	TACAACTCGTGTACAGTCACCCTACAG	M2	GAGCTGGTGTTTTGTAAACACGATG
AAGE02034767.1(1-6813)	M49	AAGAAGCATGTCCTCTGCGGTC	M50	CGAGCGAGATCAGAAATTAAGTGA

Appendix 3 Digital Appendix

Certain datasets used in this thesis have been deposited on a Centre for Genomic Research FTP server and can be accessed at:
http://cgr.liv.ac.uk:80/pbio/JTurnerThesisAppendix_17a78110d09decc1/

Contents

Chapter 2

2.1. Results of PCR screening of candidate male contigs in wild type gDNA.

2.2. The top three candidate male contig sequences in GenBank format.

2.3. Sequences of the constructs used in injection experiments in GenBank format.

Chapter 3

3.1. The assembled M locus region, BAC ends, vector sequence, and gene annotation.

3.2. RepeatMasker output for the M locus region and intron of *Nix.*

Chapter 4

4.1. The Illumina and 10x linked reads used in the analyses.

4.2. The BEDTools coverageBED data from male and female alignments.

4.3. The 10x Supernova assembly.

4.4. RepeatMasker output for the genomes of *Aedes aegypti* and *Aedes albopictus*.

4.5. Scripts used in the analyses.